Universidad Autónoma de Madrid

Facultad de Psicología

SEGURIDAD DEL BANCO DE ITEMS EN TESTS ADAPTATIVOS INFORMATIZADOS

Tesis Doctoral

Juan Ramón Barrada González

Madrid, 2009

Universidad Autónoma de Madrid

Facultad de Psicología

Departamento de Psicología Social y Metodología

SEGURIDAD DEL BANCO DE ITEMS EN TESTS ADAPTATIVOS INFORMATIZADOS

Tesis Doctoral

Autor:

Juan Ramón Barrada González

Directores:

Julio Olea Díaz

Francisco José Abad García

Índice

PARTE I – INTRODUCCIÓN

PARTE II – ESTUDIOS

PARTE I – INTRODUCCIÓN

Capítulo 1.
Test Adaptativos Informatizados. Perspectiva general.

1.1. Introducción

La Psicometría indica que los items varían en su calidad para la medición y que no todos los items son igualmente adecuados para todos los niveles de rasgo. Aquellos items que más estrechamente se vinculan con aquello que estamos midiendo presentarán una mayor correlación con el total de la escala y, al ser calibrados según algún modelo de Teoría de Respuesta al Ítem (TRI), un mayor parámetro de discriminación. Aquellos items mejor permitan evaluar a un examinado tendrán, en general, un parámetro de localización próximo al nivel de rasgo del examinado.

Teniendo en cuenta esto, si se dispone de un banco de items marcadamente más amplio que la longitud del test que deseamos administrar, resultaría conveniente ajustar los items a presentar a cada examinado. En lugar de administrar tests fijos, en los que todos los evaluados reciben idénticos items, podríamos seleccionar aquellas preguntas que en mayor medida reducen la incertidumbre sobre el nivel de rasgo de cada examinado. La selección del ítem se haría teniendo en cuenta tanto los parámetros de los items como el nivel de rasgo del examinado. Las puntuaciones de los examinados que han respondido diferentes preguntas se situarían en la misma escala mediente métodos derivados de la TRI.

El nivel de rasgo de los examinados, una variable latente, es una información que resulta inaccesible. De lo que podemos disponer es de estimaciones que esperamos que sean progresivamente más precisas según aumenta el número de items administrados. Previamente a la administración de ningún ítem, la estimación que minimiza el error es asignar a un examinado el nivel de rasgo promedio en la población. Una vez administrados uno o más items, la estimación se realizará según el patrón de respuestas a los items contestados y los parámetros de éstos.

Estas ideas básicas son las que fundamentan los Tests Adaptativos Informatizados (TAIs). Los TAIs ofrecen varias opciones atractivas, cuando se comparan con test de lápiz y papel: (a) resulta posible igualar para los diferentes niveles de rasgo la precisión de

medida; (b) resulta posible, para una misma longitud del test, mantener la precisión reduciendo a la mitad el númer de items administrados; o (c) manteniendo la longitud, se obtienen estimaciones más precisas. Las primeras propuestas teóricas son de los años setenta (p. ej., Lord, 1977; Urry, 1977), pero es en los años ochenta y noventa cuando empieza a popularizarse este modo de administración de tests. A finales del pasado siglo, ya se admistraban más de un millón de TAIs por año (Wainer, 2000a). Actualmente, importantes pruebas como el GRE (*Graduate Record Examination*), el ASVAB (*Armed Services Vocational Aptitude Battery*), el TOEFL (*Test of English as a Foreign Language*) o el GMAT (*Graduate Management Admission Test*), entre otras, son administradas de un modo adaptativo. En España, los primeros TAIs comercializados llegan años más tarde: el TRASI (Rubio & Santacreu, 2004), que mide la capacidad de razonamiento secuencial e inductivo, y eCAT (Olea, Abad, Ponsoda & Ximénez, 2004), que mide el nivel de comprensión del inglés escrito. Recientemente, se ha desarrollado un TAI desde el ámbito médico, CAT-Health (Rebollo, García-Cueto, Zardaín, Cuervo, Martínez, Alonso, Ferrer & Muñiz, 2009), para la evaluación de la calidad de vida relacionada con la salud. Hoy día, el campo de los TAIs, tanto a nivel de investigación como aplicado, y tanto a nivel nacional como internacional, es un área activa y en desarrollo.

A continuación, procederemos a: (a) ilustrar la estructura básica de un TAI; (b) definir cuáles son los objetivos que ha de satisfacer; (c) desarrollar los modos de selección de items empleados en los TAIs, puesto que es en ésta línea donde se inscribe el contenido de los estudios que presentamos.

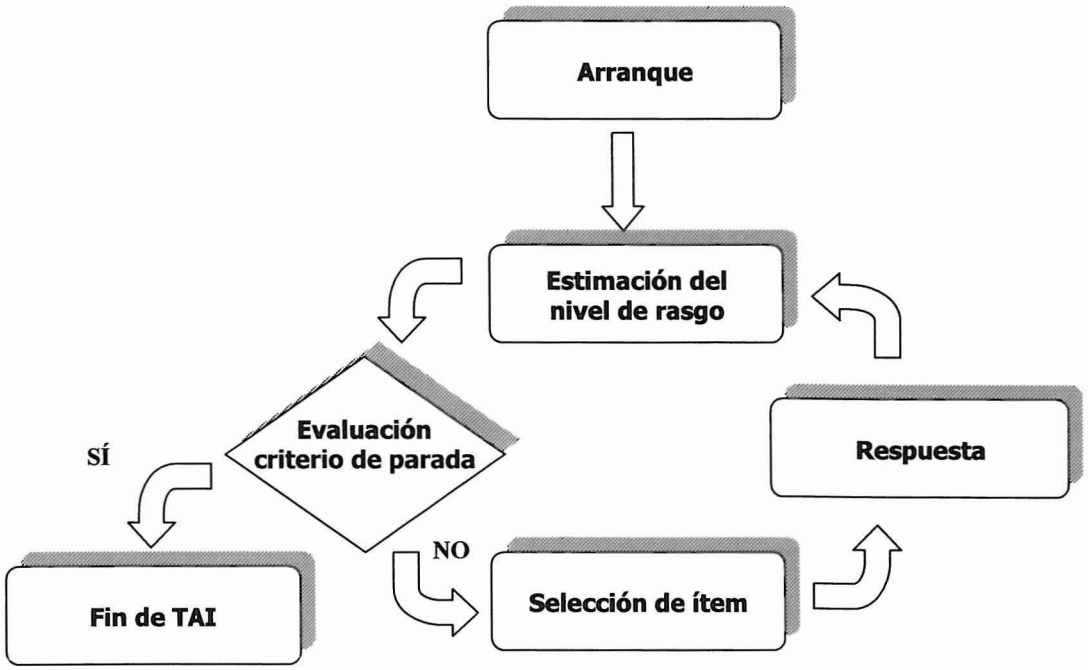
1.2. Estructura de un TAI

La Figura 1 recoge el diagrama de flujo de un TAI. En el arranque se inicializa el sistema, leyendo los enunciados y los parámetros de los items y demás requisitos del procedimiento informático. En la fase de fin del TAI se salvaría la información pertinente, se ofrecería la retroalimentación e instrucciones que correspondieran al examinado, etc. El cuerpo central del TAI lo compone un proceso iterativo con cuatro pasos. Ofrecemos una descripción de las principales propuestas para aplicar cada uno de ellos.

1.2.1. Estimación del nivel de rasgo

Hay que distinguir entre dos momentos diferentes en un TAI.

Figura 1
Diagrama de flujo de un TAI.



- El primero, cuando el examinado todavía no ha respondido a ítem alguno. En este caso, tres son las opciones que se presentan (Parshall, Spray, Kalohn & Davey, 2002): (a) asignar como nivel de rasgo el promedio muestral; (b) para incrementar la motivación de los examinados, asignar como nivel inicial un valor por debajo del promedio para aumentar la probabilidad de acierto, o (c) mediante información sobre los examinados que permita predecir sus desempeños en el test (p. ej., puntuaciones en otro test o nivel educativo) estimar un nivel de rasgo inicial diferente por examinado.
- El segundo, cuando el examinado ya ha respondido a alguna pregunta. En este caso, se aplican los métodos de estimación por máxima-verosimilitud o bayesianos para estimar el nivel de rasgo. Un supuesto básico en la mayor parte de los modelos de TRI es la independencia local. Asumiendo ésta, la probabilidad del patrón de respuestas observado es igual al producto de las probabilidades de la respuesta obtenidas en cada ítem. La función de verosimilitud es la que relaciona este producto con el continuo de niveles de rasgo:

$$L(\theta|u_1 \ldots u_n, x_1 \ldots x_n) = \prod_{i=1}^{n} \left\{ P_i(\theta)^{u_i} \left(1 - P_i(\theta)\right)^{1-u_i} \right\}^{x_i}, \quad (1)$$

donde θ es el nivel de rasgo, $P_i(\theta)$ es la probabilidad de respuesta correcta del ítem *i*, *n* es el tamaño del banco de items y x_i es el indicadores de administración / no administración (1 / 0) y u_i es el indicador de respuesta correcta / no correcta (1 / 0), ambos del ítem *i*.

El método de máxima-verosimilitud (Birnbaum, 1968) ofrece como nivel de rasgo estimado aquel en el que la función de verosimilitud encuentra su máximo:

$$\hat{\theta}_{q-1} = \arg\max_{\theta} L(\theta|u_1 \ldots u_n, x_1 \ldots x_n), \quad (2)$$

donde θ_{q-1} es el nivel de rasgo estimado después de los ***q-1*** primeros items, siendo *q* el indicador de la posición serial dentro del test del siguiente ítem a administrar. Cuando se opta por la estimación máximo-verosímil, hay que tener en cuenta que la función de verosimilitud no tiene máximo dentro de los números reales mientras el patrón de respuestas es constante, todas las respuestas aciertos o todas errores. Por eso, hasta que el patrón se rompe, las alternativas son: (a) aplicar temporalmente una estimación bayesiana; (b) fijar el rango de valores admisibles de niveles de rasgo y permitir que la estimación se sitúe en uno de los extremos; o (c) utilizar un método por escalera, mediante el cual el nivel de rasgo no es estimado sino asignado, haciendo que a respuestas correctas le sigan incrementos en el nivel de rasgo y a respuestas erróneas decrementos (Dodd, 1990).

Los métodos bayesianos incorporan información previa sobre la distribución de los niveles de rasgo en la población. De este modo, se añade el condicionante de una cierta distribución a priori de los niveles de rasgo, $g(\theta)$. La distribución a posteriori será:

$$g(\theta|u_1 \ldots u_n, x_1 \ldots x_n) = \frac{L(\theta|u_1 \ldots u_n, x_1 \ldots x_n) g(\theta)}{\int L(\theta|u_1 \ldots u_n, x_1 \ldots x_n) g(\theta) d(\theta)}. \quad (3)$$

La estimación Máximo a Posteriori (MAP; Lord, 1986; Mislevy, 1986) busca el nivel de rasgo donde es máxima esta distribución a posteriori:

$$\hat{\theta}_{q-1} = \arg\max_{\theta} g(\theta|u_1 \ldots u_n, x_1 \ldots x_n). \quad (4)$$

En el caso de una distribución a priori uniforme, MAP y máxima-verosimilitud coinciden.

En la estimación Esperanza a Posteriori (EAP; Bock & Mislevy, 1982) el nivel de rasgo estimado es el valor esperado de la distribución a posteriori:

$$\hat{\theta}_{q-1} = \int \theta g(\theta | u_1 \ldots u_n, x_1 \ldots x_n) d(\theta). \quad (5)$$

1.2.2. Evaluación del criterio de parada

Varios son los criterios de parada que pueden aplicarse en un TAI. De entre éstos, destacan las reglas de: (a) alcanzar un cierto número de items administrados; o (b) reducir la incertidumbre en la estimación del nivel de rasgo por debajo de un nivel predeterminado. Combinaciones de estos dos criterios resultan posibles. El criterio de parada a aplicar en un TAI determinado dependerá de diferentes factores. Por ejemplo, cuando multitud de especificaciones respecto al contenido del test han de ser controladas, la opción más razonable es un test de longitud fija, puesto que resulta más sencillo cumplir con los requisitos si se conoce de antemano cuántos items van a administrarse. En los casos en los que la validez aparente del test es un elemento importante también suele optarse por longitud fija.

1.2.3. Selección de items

En este punto puede distinguirse entre dos condiciones, cuando los items son seleccionados de uno en uno o cuando los items son seleccionados por bloques. En el primer caso, el test tiene tantos puntos de adaptación a la ejecución como items van a ser administrados. En el segundo caso, los puntos de adaptación se acostumbran a limitar a tres o cuatro. El término de TAI se suele reservar para el primer caso. El segundo recibe el nombre de Test Multietápico (Luecht & Nungester, 1998).

La selección de items es uno de los aspectos sobre los que más se ha investigado en el campo de los TAIs. Por ello, se describe con mayor detalle en el siguiente apartado.

1.2.4. Respuesta

Éste es el único elemento del TAI en el que interviene el evaluado. Los items que se administran en un TAI han sido calibrados bajo alguno de los modelos de TRI disponibles. Así, la probabilidad de acierto al ítem vendrá determinada por parámetros propios de los items y el nivel de rasgo del examinado. Por eso, los problemas característicos en la TRI de calibración y evaluación del ajuste del modelo son igualmente pertinentes en los TAIs, con el problemas añadido de que la matriz de respuestas de los examinados a los items

está casi vacía. Estudios relevantes en este apartado incluirían los relativos a la calibración de nuevos items insertados en un TAI operativo (Ban, Hanson, Wang, Yi & Harris, 2001; Ban, Hanson, Yi & Harris, 2002; Chang & Lu, en prensa), el estudio del funcionamiento diferencial (Lei, Chen & Yu, 2006, Zwick, 2000) o el estudio de patrones anómalos (Nering, 1997).

1.3. Selección de items en TAIs

En general, la selección de items en un TAI se realiza definiendo un subconjunto del banco de items y buscando cuál es el ítem de este sub-banco que optimiza una determina función de valoración. A esta base general se pueden añadir otras restricciones. El modo de determinar ese subconjunto de preguntas y la función de valoración a emplear vendrá determinado por el peso relativo de los diferentes objetivos que ha de satisfacer el test. Por ello, será desarrollando los objetivos de TAI cómo se presentarán las principales reglas de selección de items que se han ofrecido.

Davey y Parshall (1995) identifican tres objetivos básicos a cumplir mediante un TAI. A éstos añadiremos un cuarto.

1. Permitir la estimación precisa del nivel de rasgo de los evaluados.
2. Limitar la probabilidad e implicaciones de una filtración de items.
3. Garantizar el ajuste a las especificaciones de contenido de la prueba.
4. Facilitar el mantenimiento del banco de items.

En los siguientes apartados, desarrollaremos estos puntos más extensamente.

1.3.1. Permitir la estimación precisa del nivel de rasgo de los evaluados

Al igual que con el resto de tests, la interpretación de las puntuaciones de un TAI puede estar orientada a criterio o a norma. En el primer caso, lo que se busca es clasificar a los examinados en una o más categorías (apto o no apto; nivel bajo, medio o alto). Lo relevante no es la precisión en la estimación del nivel de rasgo, sino la precisión y consistencia de las clasificaciones. Dos son las vías básicas para clasificar examinados. La primera, conseguir estimaciones precisas de su nivel de rasgo y comparar éstas con los puntos de corte. La segunda supone centrarse únicamente en si el examinado está por encima o por debajo del punto de corte. Estas dos opciones dan lugar a diferentes modos de seleccionar items (Thompson, 2009).

El problema de interpretación referida a norma supone objetivos y, por tanto, criterios diferentes para la selección de items. En este caso, buscamos situar a todos los

examinados en un continuo con la mayor precisión posible. Nos detendremos en los modos que se han propuesto para determinar qué items son los más apropiados para este objetivo.

La mayor parte de las reglas propuestas seleccionan el ítem que optimiza una cierta función de valoración. Esta función puede tomar como entrada un único valor, el nivel de rasgo estimado, o un intervalo de rasgos. La función más comúnmente empleada es la función de información de Fisher evaluada únicamente para el rasgo estimado.

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}, \tag{6}$$

donde $P_i'(\theta)$ es la primera derivada de $P_i(\theta)$.

El ítem seleccionado es aquel con máximo resultado (Lord, 1980):

$$j = \arg\max_{i \in B_q} I_i(\hat{\theta}), \tag{7}$$

donde B_q define aquellos items del banco que son evaluados para determinar cuál será administrado en la q-ésima posición. A este criterio lo llamaremos selección por máxima información puntual.

La base de este criterio es una propiedad asintótica de la estimación máximo-verosímil. Bajo ciertas condiciones, el recíproco de la información de Fisher para el nivel estimado equivale al error típico de medida (Chang & Ying, 2009). Por tanto, incrementos en la información reducirán el error.

Esta función de valoración no está exenta de problemas, tanto en términos de precisión como de seguridad. Empezaremos por las limitaciones en la precisión, dejando las de seguridad para una sección posterior. Respecto a la precisión, el inconveniente principal viene causado por seleccionar items considerando el nivel estimado como idéntico al nivel real, obviando el error de medida, especialmente presente cuando son pocos los items administrados. Un modo de tener en cuenta la incertidumbre sobre la ubicación del nivel de rasgo es ponderar la función de información de Fisher mediante la función de verosimilitud (Veerkamp & Berger, 1997). De este modo, para los primeros items, cuando la función de verosimilitud es más plana, se buscarán items que ofrezcan información para todo el rango posible de niveles de rasgo, en lugar de información concentrada para un nivel estimado que puede distar en gran medida del rasgo estimado final:

$$j = \arg\max_{i \in B_q} \int I_i(\theta) L(\theta | u_1 \ldots u_n, x_1 \ldots x_n) d(\theta). \tag{8}$$

La función de información de Fisher indica la capacidad de un ítem para discriminar entre valores adyacentes de rasgo. Cuando la estimación es pobre, resultaría más oportuna una función que nos permitiera discriminar entre cualesquiera pares de valores. Esto lo permite la función Kullback-Leibler, propuesta como función de valoración en los TAIs por Chang y Ying (1996). Lo recomendable es ponderar la función Kullback-Leibler por la función de verosimilitud (Barrada, Olea, Ponsoda & Abad, 2009). Esta función de valoración se expresa del siguiente modo:

$$j = \arg\max_{i \in B_q} \int KL_i(\theta \,||\, \hat{\theta}) L(\theta | u_1 \ldots u_n, x_1 \ldots x_n) d(\theta), \tag{9}$$

donde

$$KL_i(\theta \,||\, \hat{\theta}) = P_i(\hat{\theta}) \ln\left[\frac{P_i(\hat{\theta})}{P_i(\theta)}\right] + [1 - P_i(\hat{\theta})] \ln\left[\frac{1 - P_i(\hat{\theta})}{1 - P_i(\theta)}\right]. \tag{10}$$

Van der Linden (1998) sugiere otra aproximación, las reglas de selección bayesianas. Desde su punto de vista, no tiene sentido buscar el ítem informativo en el nivel de rasgo estimado, puesto que una vez administrada una nueva pregunta la estimación cambiará. Estos cambios entre estimaciones sucesivas serán mayores cuanto más próximos al comienzo del test nos encontremos (Chang & Ying, 2008). La idea de van der Linden es buscar el ítem que maximice la información en los niveles de rasgo donde se situaría la estimación en el caso de acierto o error, ponderando por la probabilidad de acierto o error. Para ello, ofrece varias propuestas. A modo de ejemplo, la siguiente:

$$j = \arg\max_{i \in B_q} \{I_i(\hat{\theta}_{u_i=1}) P_i(\hat{\theta}) + I_i(\hat{\theta}_{u_i=0}) [1 - P_i(\hat{\theta})]\}, \tag{11}$$

donde $\hat{\theta}_{u_i=1}$ y $\hat{\theta}_{u_i=0}$ serían los niveles de rasgo estimados en el caso de respuesta correcta o incorrecta al ítem *i*, respectivamente.

Las funciones de valoración alternativas a la de máxima información de Fisher comparten varias características: (a) a medida que aumenta el número de items administrados, convergen hacia la seleccion de items por máxima información puntual (Ecuacion 7), puesto que la función de verosimilitud es más apuntada y los niveles de rasgo en caso de acierto o fallo están cada vez más próximos entre sí); (b) por tanto, a mayor número de items, menor es el beneficio en precisión por el uso de las funciones alternativas en comparación con la selección por máxima información puntual (Chen, Ankenmann & Chang, 2000; Chen & Ankenmann, 2004).

1.3.2. Limitar la probabilidad e implicaciones de una filtración de items

Mientras que los examinadores mantienen para todo tipo de test la voluntad de conseguir una estimación precisa del nivel de rasgo de los examinados, el objetivo de los evaluados puede variar según las consecuencias del resultado de la evaluación (Wainer, 2000b). En las situaciones en las que un elevado error de estimación les suponga consecuencias adversas, los examinados desearán un nivel estimado tan próximo a su nivel real como resulte posible (por ejemplo, en ciertas pruebas de diagnóstico psicopatológico donde tanto las falsas alarmas como las omisiones pueden conllevar efectos negativos para los examinados). Pero en otros casos lo más ventajoso para los examinados es conseguir un cierto nivel de rasgo, con independencia de si éste se corresponde o no con el real. Por ejemplo, en una prueba de reválida de Bachillerato en la que no superarla supone la no homologación de varios años de estudio. En este caso, la mayor parte de los evaluados deseará un resultado de apto con independencia de su nivel real.

En los procesos de evaluación en los que los objetivos de examinadores y examinados no concuerdan es probable que una parte de los evaluados busquen el modo de incrementar artificialmente sus calificaciones. En una revisión de prácticas tramposas en exámenes, Cizek (1999) señaló que la mitad o más de los estudiantes admiten copiar. Varias son las opciones para esto (Davey & Nering, 2002). Por ejemplo, mirar alrededor buscando algún examinado con una pregunta idéntica a la nuestra y copiarle la respuesta o llevar auriculares y micrófonos ocultos para que alguien vaya enviando desde el exterior las respuestas. Una alternativa especialmente provechosa es llevar de antemano conocido parte del banco de items, de tal modo que la probabilidad de respuesta correcta a esas preguntas sea elevada, con independencia del parámetro de rasgo. Como señal de lo útil de esta estrategia, un 76% de los estudiantes universitarios preguntados por Stern y Havlick (1986) reconocieron preguntar a estudiantes ya evaluados por el contenido de su examen.

El riesgo de conocimiento previo de items está especialmente presente en los TAIs (Chang, 2004). Esto se debe a dos motivos básicos:

- Los TAIs cobran sentido cuando se ofrecen de forma continuada. A diferencia de algunos programas en los que únicamente se administran las pruebas unas pocas ocasiones al año, la mayor parte de los TAIs permanecen activos todo el año. En las pruebas que no se ofrecen de un modo continuo, los items empleados son, por lo general, automáticamente descartados para ocasiones posteriores. En los TAIs, los

bancos de items son relativamente estáticos a lo largo del tiempo. Esto supone que un examinado puede preguntarle a personas previamente evaluadas por los items que recibieron. También puede recurrir a academias de preparación de exámenes o sitios web especializados en el filtrado de preguntas.

- Conocer ciertos items de antemano será provechoso si algunos de estos le son presentados al examinado. La mayor parte de las reglas de selección implican una alta probabilidad de que haya coincidencia entre los items preconocidos y aquellos administrados. Por un lado, al comienzo del test los niveles de rasgo estimados tienden a ser muy parecidos para los diferentes examinados. En general, a mismo nivel de rasgo estimado, mismo ítem administrado. Por otro lado, puesto que las reglas de selección de items buscan aquellos items con mejores propiedades métricas, los items seleccionados se concentran entre aquellos de mayor parámetro de discriminación (Barrada, Olea et al., 2009; Li & Schafer, 2005).

La consecuencia de esto es una elevada coincidencia entre los items recibidos por diferentes examinados. Definiremos la tasa de solapamiento como la proporción de items que dos examinados tomados al azar comparten (Way, 1998). Chen, Ankenmann y Spray (2003) han mostrado que la tasa de solapamiento, para tests de igual longitud e igual tamaño del banco de items, crece linealmente con la varianza de las tasas de exposición de los items. La tasa de solapamiento puede calcularse mediante la siguiente ecuación:

$$T = \frac{n}{Q} S^2_{P(A)} + \frac{Q}{n}, \tag{12}$$

donde Q es la longitud del test y $S^2_{P(A)}$ es la varianza de las tasas de exposición.

La tasa de solapamiento y la distribución de las tasas de exposición se han convertido en las dos variables más habituales para evaluar el riesgo de conocimiento previo (Chang & Zhang, 2002). Se ha entendido que a mayor tasa de solapamiento, más provechoso puede ser el preconocimiento de items. Igualmente, se ha considerado que distribuciones más homogéneas de tasas de exposición son preferibles.

Varias son las propuestas que se han formulado hasta el momento para limitar este riesgo.

1.3.2.1. Restricciones en el banco presentable

Entendemos por banco presentable (B_q) el subconjunto de preguntas del banco de items que son evaluadas por la regla de valoración a la hora de seleccionar el ítem para su administración en la posición q. La restricción mínima a la hora de configurar B_q es no incluir aquellos items administrados al examinado en posiciones previas del test. Una parte importante de las propuestas para mejorar la seguridad del banco parten de limitar el tamaño de B_q, de tal modo que se reduzca la probabilidad de que aquellos items con tendencia a ser más expuestos formen parte de él. Cuatro son los modos básicos de hacer esto:

- *Restricción de tasa máxima de exposición:*

  Una opción para homogeneizar las tasas de exposición de los items es limitar la proporción de examinados que pueden recibir los items. Esta tasa máxima, fijada de antemano por la institución o empresa responsable del test, la llamaremos r^{max}. El modo de limitar la tasa máxima es reducir la frecuencia con la que los items con tasas de exposición elevadas entran en B_q (Barrada, Abad & Veldkamp, 2009).

  El primer método propuesto para esto es el método Sympson-Hetter (Hetter & Sympson, 1997; Sympson & Hetter, 1985), en el que se definen dos eventos diferentes: (a) E_i, que el ítem i sea marcado como elegible (incorporado a B_q); y (b) A_i, que el ítem i sea administrado. Puesto que cualquier ítem administrado ha de ser elegible, se cumple que:

$$P(A_i) = P(A_i \mid E_i)P(E_i). \tag{13}$$

  $P(A_i)$, la probabilidad de administración o tasa de exposición, es el valor que quiere controlarse. El modo de conseguirlo es fijando convenientemente valores de $P(E_i)$, la probabilidad de que un ítem entre en B_q. Estas probabilidades de elegibilidad se calculan mediante un proceso iterativo. Los parámetros $P(E_i)$ para el ciclo $t+1$ derivan de hacer $P(A_i)$ igual a r^{max} en la ecuacion anterior y fijar en 1 el valor máximo de estos parámetros:

$$P^{(t+1)}(E_i) = \begin{cases} 1 & si \quad P^{(t)}(A_i)/P^{(t)}(E_i) \le r^{\max} \\ r^{\max} P^{(t)}(E_i)/P^{(t)}(A_i) & si \quad P^{(t)}(A_i)/P^{(t)}(E_i) > r^{\max} \end{cases}. \tag{14}$$

  Cuando el valor de la tasa máxima de exposición se estabiliza o cuando se alcanza un número predefinido de ciclos, el proceso de simulación para establecer los parámetros

$P(E_i)$ finaliza. Una vez establecidos estos valores, para cada examinado se generan tantos números aleatorios dentro del intervalo uniforme (0, 1) como items componen el banco. Sólo en el caso de que el número aleatorio sea menor a $P(E_i)$ el ítem *i* formará parte de B_q.

Este método presenta varios problemas. Primero, no todas las tasas quedan por debajo del límite (van der Linden, 2003). Segundo, el método funcionará en la medida en la que la distribución de rasgo de la simulaciones coincidan con la distribución de rasgo de los examinados (Chen & Doong, 2008). Tercero, las simulaciones han de ser repetidas cada vez que se incorpora o se retira cualquier ítem del banco (Chang & Harris, 2002). Cuarto, las simulaciones necesarias consumen tiempo, si bien se han propuesto varias vías para reducirlo (Barrada, Olea & Ponsoda, 2007; Chen & Doong, 2008; van der Linden, 2003).

Otra alternativa es el método restringido de Revuelta y Ponsoda (1998). Este método, a diferencia del método Sympson-Hetter, no requiere simulaciones previas, sino que ajusta la pertenencia o no de cada ítem a B_q para cada nuevo examinado, según las tasas de exposición encontradas. Aquellos items con una tasa de exposición mayores de r^{max} hasta un cierto examinado son retirados de B_q y no vuelven a ser incorporados hasta que su tasa se sitúa por debajo del límite. Con este método, la pertenencia o no a B_q es determinista, no probabilística como en el método Sympson-Hetter, y se ajusta para cada nuevo examinado. Siendo *f* el indicador de la posición ordinal del examinado dentro del total de examinados evaluados, los valores de $P(E_i)$ se ajustan mediante la siguiente fórmula:

$$P^{(f+1)}(E_i) = \begin{cases} 1 & si \quad P^{(1..f)}(A_i) < r^{\max} \\ 0 & si \quad P^{(1..f)}(A_i) \geq r^{\max} \end{cases}. \tag{15}$$

En el método de elegibilidad del ítem (van der Linden & Veldkamp, 2004) la pertenencia a B_q no es determinista. Este sistema, al igual que el método restringido, no requiere de simulaciones previas y la probabilidad de entrar en B_q no es fija para todos los examinados, sino que se va adaptando según la administración o no del ítem a examinados previos:

$$P^{(f+1)}(E_i) = \begin{cases} 1 & si \quad P^{(1..f)}(A_i)/P^{(f)}(E_i) \leq r^{\max} \\ r^{\max} P^{(f)}(E_i)/P^{(1..f)}(A_i) & si \quad P^{(1..f)}(A_i)/P^{(f)}(E_i) > r^{\max} \end{cases}. \tag{16}$$

De entre las tres propuestas descritas, la alternativa ofrecida por van der Linden y Veldkamp (2004) es la que parecer ser preferible (Barrada, Abad & Veldkamp, 2009): (a) no necesita simulaciones previas para calcular los parámetros de control de la exposición; (b) es independiente de supuestos sobre la distribución del nivel de rasgo de los examinados; y (c) satisface de un modo casi completo la restricción de que no haya ninguna tasa de exposición mayor a r^{max}.

Un control de la tasa máxima tal y como se ha planteado deja la opción de sobreexposición de ciertos items y alto solapamiento cuando se condicionada a niveles de rasgo (Davey & Parshall, 1995). Por ejemplo, sería posible que casi todos los examinados con alto nivel de rasgo, infrecuentes en la población, recibieran idénticos items y, aún así, la tasa de exposición de éstos estuviera por debajo de r^{max}. Por eso, se ha propuesto controlar la tasa máxima de exposición condicionada a nivel de rasgo, tanto real (Stocking & Lewis, 1998, para una variante del método Sympson-Hetter) como estimado (Stocking & Lewis, 2000, para el mismo método; van der Linden & Veldkamp, 2007, para el método de elegibilidad del ítem). En este caso, a cada ítem ya no le corresponde un único parámetro de control de la exposición, sino tantos parámetros como intervalos en los que hayamos dividido el continuo del nivel de rasgo.

- *Restricciones de tasa máxima de exposición y tasa de solapamiento:*

  Distintas funciones de valoración de items, con la misma restricción de la tasa máxima, no conllevan la misma tasa de solapamiento. No existe una función que determine para cierto nivel de r^{max} qué tasa de solapamiento producirá.

  Chen y Lei (2005) ampliaron el método Sympson-Hetter para permitir el control simultáneo de r^{max} y la tasa de solapamiento. Este método se basa en la relación conocida entre varianza de las tasas de exposición y solapamiento (Ecuación 12). Fijando la tasa de solapamiento objetivo (T^{obj}), es posible calcular la varianza de las tasas de exposición que llevaría a T^{obj}. Esta varianza la llamaremos S^{obj}. En cada nueva iteración del ciclo de simulaciones requerido para calcular los parámetros de control de la exposición, la tasa de exposición objetivo para los items en el ciclo $t+1$ se sitúa en:

$$P^{(t+1)}(A_i) = S^{obj}\left(\frac{P^{(t)}(A_i) - \frac{Q}{n}}{S_{(t)}}\right) + \frac{Q}{n}, \quad (17)$$

donde $S_{(t)}$ es la desviación típica de las tasas de exposición para el ciclo t. De este modo, se fija la distancia en desviaciones típicas con respecto a la tasa media (Q/n) para los diferentes items de ciclo en ciclo. El proceso de ciclos de simulación permite ir afinando los parámetros de control de la exposicíon hasta conseguir el control de r^{max} y T^{obj}.

Desarrollos posteriores de esta idea han permitido que este control simultáneo se pueda hacer sin simulaciones previas para conseguir las probabilidades de control de la exposición (Chen, Lei & Liao, 2008).

- *Bancos rotatorios:*

  Algunos programas de evaluación cuentan con bancos de miles de items. Esto permite construir diferentes sub-bancos de menor tamaño, todos ellos con de la misma capacidad de medida para los diferentes niveles de rasgo (Mills & Steffen, 2000). Hay varias estrategias para manejar los sub-bancos: (a) pueden irse alternando por emplazamientos de evaluación, en aquellos programas en los que la evaluación se realiza desde un número limitado de centros; (b) pueden irse activando y desactivando según una programación temporal; o (c) al comenzar el test, puede asignarse aleatoriamente a cada examinado el banco que le corresponderá. De este modo, se limita la tasa máxima de exposición de los items (un ítem no puede tener mayor tasa que la proporción de ocasiones en las que se emplea el sub-banco o sub-bancos a los que pertenece).

  Dos son los diseños de bancos rotatorios, con coincidencia de items o sin ella (Ariel, Veldkamp & van der Linden, 2004). Los bancos sin coincidencia son aquellos en los que la extracción de items del banco maestro (aquel que contiene todos los items) se realiza sin reposición, de tal modo que los sub-bancos no comparten ítem alguno. Los bancos con coincidencias permiten que aquellos items con menor tasa de exposición formen parte de varios sub-bancos.

  Cuando se han comparado la estrategia de bancos rotatorios y de restricción de r^{max}, se ha encontrado que sus resultados en precisión y seguridad son casi indistinguibles

(Barrada, Olea & Abad, 2008), con el método de los bancos rotatorios superando ligeramente a la opción del control de tasas máximas.

- *Banco presentable variable según posición del ítem en el test:*

  Los métodos de restricción de r^{max} (con o sin control de la tasa de solapamiento) afrontan el problema de la seguridad del banco reduciendo la sobreexposición de los items más populares. Restricciones de r^{max} incrementan la exposición de aquellos items ligeramente por debajo en calidad métrica de aquellos cuya exposición se limita. Este modo de actuar tiene un efecto más bien limitado a la hora de incrementar la tasa de exposición de los items nunca o apenas utilizados.

  Los métodos estratificados son una propuesta para incrementar el uso de aquellos items infraexpuestos (Chang & Ying, 1999). En éstos, B_q se hace variable según la posición del ítem dentro de la secuencia de preguntas en el test. Al comienzo del test, cuando el error de medida es máximo y la estimación del nivel de rasgo es más inestable, B_q está compuesto únicamente por aquellos items con menor capacidad discriminativa. A medida que el test avanza, B_q se va componiendo por items de mayor calidad. De este modo: (a) se fuerza que items que no serían nunca empleados lo sean, ya que son los únicos disponibles; y (b) se reservan los items de mayor discriminación para las fases finales del test, cuando la estimación es más precisa.

1.3.2.2. Cambios en la función de valoración de items

Las funciones de valoración que hemos revisado buscan maximizar la precisión de medida. Otras funciones, sin embargo, han sido desarrolladas con la idea de incrementar la seguridad del banco de items. La idea básica para ello es reducir la sobre-exposición de los items con alto parámetro de discriminación e incrementar las tasas de exposición de aquellas preguntas que, con las reglas orientadas a la precisión, nunca o casi nunca son presentadas. De las funciones de valoración propuestas con este fin, destacamos:

- *Los métodos de distancia con respecto a la dificultad del ítem:*

  Una opción de romper la relación entre parámetro de discriminación y tasa de exposición es seleccionar los items atendiendo únicamente a su parámetro de localización, sin tener en cuenta su capacidad discriminativa. Un ítem calibrado según un modelo dicotómico ofrece información máxima para valores de rasgo iguales al parámetro de localización (modelo de 1 y 2 parámetros) o ligeramente por encima de éste (modelo de 3 parámetros). En el modelo de 3 parámetros, el nivel de rasgo en el

que un ítem alcanza su información de Fisher máxima se sitúa en (Hambleton & Swaminathan, 1985):

$$\theta_i^{\max} = b_i + \frac{\ln\left[1+(1+8c_i)^{1/2}\right]-\ln(2)}{1.7a_i}. \quad (18)$$

Una posible función de valoración sería, pues, seleccionar los items con distancia mínima entre el nivel de rasgo estimado y su parámetro *b*:

$$j = \arg\min_{i\in B_q} \left|\hat{\theta} - b_i\right|. \quad (19)$$

O entre el nivel de rasgo estimado y el nivel de rasgo en el que se obtiene la máxima información del ítem:

$$j = \arg\min_{i\in B_q} \left|\hat{\theta} - \theta_i^{\max}\right|. \quad (20)$$

De este modo, estaremos: (a) equilibrando en gran medida las tasas de exposición de los items (Li & Schafer, 2005); y (b) administrado los items a aquellos examinados para los que resultan más convenientes. El problema es una reducción en la precisión de medida, al considerar equivalentes en la selección items de diferente capacidad discriminativa.

Por eso, los métodos de distancia con respecto a la dificultad suelen a aplicarse combinados con la estrategia de banco presentable variable según posición del ítem en el test. El método alfa-estratificado (Chang & Ying, 1999) se ajusta a este perfil y ha sido, probablemente, la propuesta que ha recibido más atención en los últimos años en la investigación de la seguridad en TAIs. En este método, la capacidad de discriminación de los items aumenta según avanza el test y la selección se realiza teniendo en cuenta únicamente el rasgo estimado y el parámetro *b*. Barrada, Mazuela y Olea (2006) propusieron el equivalente al método alfa-estratificado cuando se incluye el cambio en la ubicación del nivel de máxima información que introduce el parámetro de pseudo-azar en el modelo de 3 parámetros. Igualmente, propusieron estratificar el banco no según el parámetro *a*, sino según la información máxima que puede alcanzar un ítem (Hambleton & Swaminathan, 1985):

$$I_i^{\max} = \frac{1.7^2 a_i^2}{8\left(1-c_i^2\right)}\left[1-20c_i-8c_i^2+(1+8c_i)^{3/2}\right]. \quad (21)$$

- *El método progresivo:*

  Esta propuesta, de Revuelta y Ponsoda (1998), supone empezar el test con selección

aleatoria de items y, según la prueba va avanzando, ir incrementando el peso de la información en la selección de items. Este modo de proceder permite reducir la infraexposición (ningún ítem tiene una tasa de exposición nula, ya que el test comienza con selección al azar) y también reducir ligeramente la sobreexposición (parte del problema de la sobreexposición proviene del limitado número de posibles niveles de rasgo estimados al comienzo del test, lo que dispara la demanda de items para estos pocos niveles). Esto se consigue con una perdida nula o muy pequeña en la precisión de medida.

Concretamente, en el método progresivo la función se valoración se compone de dos elementos, uno aleatorio (R_i) dentro del intervalo [0, $\max_{i \in B_q} I_i(\hat{\theta})$] y el otro, la información de Fisher del ítem:

$$j = \arg\max_{i \in B_q} \left[(1 - W_q) R_i + W_q I_i(\hat{\theta}) \right]. \quad (22)$$

El peso del componente de la información (W_q) varía según la posición del ítem en el test. Revuelta y Ponsoda (1998) propusieron para definir W_q:

$$W_q = \frac{q-1}{Q}, \quad (23)$$

donde q es la posición serial del ítem y Q la longitud del test. Esta ecuación supone una reducción lineal del componente aleatorio, empezando por selección aleatoria y acabando por selección casi únicamente determinada por la información de Fisher.

Varias son las razones que explican cómo es posible administrar items al azar y que esto no incremente el error de medida:

- Los items administrados por el método de máxima información puntual van siendo cada vez menos informativos, mientras que con el método progresivo el patrón es el inverso, haciendo que la información acumulada al final del test sea equivalente.
- El método de máxima información puntual presenta problemas cuando, al comienzo del test, examinados de alto nivel de rasgo fallan un ítem o examinados de rasgo bajo lo aciertan (Rulison & Loken, 2009). En tales casos, el nivel estimado se desplaza hacia el extremo incorrecto y, puesto que los items administrados son de alto parámetro *a*, la función de verosimilitud es muy apuntada. Así las cosas, es necesaria la presentación de una gran cantidad de items hasta conseguir que el nivel estimado se aproxime al nivel real. El

método progresivo, dada la importancia de la selección aleatoria, reduce el impacto de estos patrones anómalos al comienzo del test.

- *Métodos mixtos:*

  No hay razón alguna por la que haya que mantener constante la función de valoración de items a lo largo de todo el test. Distintas propuestas se han planteado de combinaciones de funciones, todas ellas compartiendo la idea de empezar por métodos menos precisos pero más seguros y, según el test avanza, pasar a funciones donde la información desempeña un papel más importante. Li y Schafer (2005) proponen empezar por selección aleatoria, pasar a selección según distancia a dificultad y acabar por máxima información. Leung, Chang y Hau (2005) y Barrada, Abad y Olea (2008) optan por estratificar el banco, empezar por selección según distancia a dificultad y acabar, también, por máxima información puntual.

Como se puede apreciar, hay una amplia variedad de métodos disponibles para la mejora de la seguridad del banco en TAIs (Georgiadou, Triantafillou & Economides, 2007). El patrón de resultados habitualmente encontrado señala la existencia de un balance entre precisión y seguridad, de tal modo que incrementos en un objetivo supone decrementos en el otro (Chang & Ansley, 2003; Finkelman, Nering & Roussos, 2009; Stocking & Lewis, 2000). Esta idea del balance será explorada con mayor detalle en varios de los estudios de la tesis.

1.3.3. Restricciones de contenido de los tests

Items pertenecientes a una misma dimensión pueden variar en los subdominios que cubren (p. ej., en un test de matemáticas puede haber items de aritmética, probabilidades y trigonometría). Previamente a la puesta en funcionamiento del programa de evaluación, la agencia encargada del test desarrolla una tabla de especificaciones en la que se detalla el rango de items que cada examinado ha de responder por subdominio.

La agencia responsable del test puede desear controlar otra multitud de aspectos de los items. Por ejemplo, impedir la presencia simultánea para un examinado en su test de 'items-enemigos', aquellos que, de presentarse uno, no ha de administrarse el otro. En algunos casos se controla la proporción de ocasiones en las que la respuesta correcta corresponde a cada una de las alternativas de respuesta. La importancia dentro del test de enunciados referidos a distintos sexos o diferentes etnias también puede ser un elemento a controlar. Algunos programas de evaluación buscan igualar los tests de los

diferentes examinados en número de palabras o en tiempo de evaluación. Como se ve, la amplitud y variedad de las restricciones a imponer en un TAI puede ser muy amplia.

Varias han sido las propuestas formuladas para conseguir cumplir con las restricciones formuladas. Nos centraremos en las cuatro principales:

- *Métodos de espiralización:*

  Estos métodos, inicialmente propuestos por Kingsbury y Zara (1991), sirven únicamente cuando la restricción es respecto a un único atributo categórico y cuando la especificaciones definen un único valor como admisible, no un rango de valores. En la propuesta original, para cada categoría y previamente a la selección de cada ítem, se calcula la división del número de items administrados entre el número de items a administrar por categoría. El ítem a presentar será seleccionado entre los pertenecientes a la categoría con menor resultado en la división, esto es, máxima discrepancia entre el estado actual y el objetivo. Una variante es utilizar el resultado de esas divisiones para construir una distribución multinomial y asignar aleatoriamente la categoría de la que se escogerá el ítem (Chen & Ankenmann, 2004). El inconveniente básico de este método es lo limitado de los casos en los que se puede aplicar. Su mayor ventaja, la simplicidad.

- *Modelo de desviación ponderada*:

  En esta propuesta de Stocking y Swanson (1993) cada uno de los atributos a controlar recibe una ponderación determinada por expertos. Los límites especificados en el diseño del test dejan de ser considerados como objetivos estrictos y pasan a ser objetivos deseables. La precisión en la medida de los examinados es considerada también un objetivo con su correspondiente peso. Los items seleccionados son aquellos que minimizan la desviación entre los objetivos del test y lo obtenido en el caso de administrar tal ítem:

$$j = \arg\min_{i \in B_q} \sum_{h=1}^{H} z_h \left| \pi_{i,h} - \gamma_h \right|, \qquad (24)$$

  donde H es el número de restricciones incluidas en el TAI, z_h es el peso asignado a cada una de estas restricciones, $\pi_{i,h}$ es el valor para la restricción h en el caso de ser administrado el ítem i y γ_h es el objetivo para la restricción h.

  Este procedimiento requiere que se determinen los pesos por objetivo (con un cierto

componente de ensayo y error) y no garantiza que se satisfagan las especificaciones por completo.

- *Aproximación del test en la sombra*:

  La aproximación del test en la sombra (van der Linden, 2000; van der Linden & Reese, 1998), basada en métodos de programación lineal, construye para cada nuevo ítem a seleccionar un test completo tal que: (a) se satisfagan todas las restricciones; (b) contenga todos aquellos items ya administrados; y (c) sea el test óptimo según la regla de selección. El ítem administrado será aquel que, formando parte del test en la sombra y no habiendo sido administrado al examinado, resulta óptimo según la regla de selección. El test resultante cumple por completo las restricciones y es el más adecuado desde el punto de vista de la regla de selección. La principal diferencia con respecto a otros métodos de restricción de contenidos es que, mientras que en éstos los items se seleccionan de uno en uno, el método del test en la sombra construye, para cada nuevo ítem a administrar, un test completo. Esto garantiza que el test administrado sea óptimo.

- *Método del índice de máxima prioridad*:

  Ésta es la propuesta más reciente (Cheng & Chang, 2009; Cheng, Chang, Douglas & Guo, 2009). Antes de la selección de un ítem, se calcula la 'cuota restante' ($\tau_{i,h}$) para cada restricción, que es la proporción de items que falta por administrar correspondientes a esa especificación. Los valores de $\tau_{i,h}$ para aquellas restricciones que el ítem no puede cubrir son fijados a 1. Para cada uno de los items que componen el banco, se calcula el producto de las cuotas restantes y este valor, a su vez, se multiplica por la función de valoración del ítem según la regla de selección (V_i). La pregunta con resultado máximo tras estas operaciones es la administrada, dado que ofrece simultáneamente mejor resultado combinado en la función de valoración y en la satisfacción de las restricciones:

$$j = \arg\max_{i \in B_q} V_i \prod_{h=1}^{H} \tau_{i,h} \,. \tag{25}$$

  Al tratarse de un producto, una vez que un objetivo ha sido satisfecho ningún ítem adicional que cubra tal requisito puede ser administrado. Por esto, el método del índice de máxima prioridad supone una satisfacción plena de los restricciones del test.

El método de espiralización es sencillo, pero limitado en su aplicabilidad. El modelo de desviación ponderada ha sido durante años el método preferido por las empresas y agencias responsables de los TAIs, si bien la selección secuencial de items no es la más eficaz y, como hemos dicho, puede llevar a que no se cumplan las especificaciones. La aproximación del test en la sombra es versátil, la selección simultánea de items es superior a la secuencial y el método es capaz de incorporar tantas restricciones como resulten necesarias sin necesidad de asignarles pesos. Cuando se ha comparado la aproximación del test en la sombra con el modelo de desviación ponderada se ha encontrado que ambos presentan resultados equivalentes en precisión de medida, si bien la primera opción ofrece un mejor control de contenidos (van der Linden, 2005). El problema del test en la sombra es su mayor complejidad, tanto matemática como de programación. El método del índice de máxima prioridad, que empieza a estudiarse recientemente, es eficaz y sencillo para la aplicación de las restricciones, si bien supone una pérdida en la precisión de medida. Al comparar el método del índice de máxima prioridad con el modelo de desviación ponderada, Cheng y Chang (2009) encontraron que ambos métodos eran equivalentes en precisión, si bien el segundo se ajustaba mejor a las especificaciones del test. Ahora bien, el estudio de Cheng y Chang hay que tomarlo con cautela, puesto que para el método de desviación ponderada no incluyeron restricción de tasa máxima de exposición, mientras que sí que lo hicieron para el método del índice de máxima prioridad, por lo que la comparación se realiza sobre métodos que difieren en un aspecto clave en el diseño del TAI. Por el momento, la propuesta de van der Linden y la de Cheng no han sido comparadas. Es probable que su funcionamiento relativo dependa de la cantidad y complejidad de las restricciones a incorporar. Mientras que la aproximación del test en la sombra ha demostrado poder satisfacer 400 especificaciones (van der Linden, 2005), el método del índice de máxima prioridad ha sido puesto a prueba con únicamente 21 restricciones (Cheng & Chang, 2009).

1.3.4. Facilitar el mantenimiento del banco de items

Todo banco de items requiere de un cierto mantenimiento (Mills & Stocking, 1996). Con el tiempo, el contenido de los constructos puede variar, haciéndose necesario el diseño de preguntas nuevas y la supresión de algunas antiguas. En algunos países existe legislación que obliga a hacer accesible al público parte del contenido del banco, para que futuros examinados reduzcan su incertidumbre sobre el contenido de la prueba. Al así hacerlo, estos items pasan a ser inservibles y han de desarrollarse y calibrarse otros que

los reemplacen. Aspectos relativos a la seguridad también invitan a la retirada de items. Aquellos que han sido utilizados más allá de un cierto límite (sea este límite un tiempo de pertenencia al banco o un número de examinados que han recibido la pregunta) son suprimidos y, si no deseamos reducir el tamaño del banco, han de ser reemplazados por otros. El coste de cada nuevo ítem dependerá de multitud de factores. Según Buyske (2005), puede fácilmente superar los 100 dólares por pregunta. Luecht (2005) lo sitúa más allá, desde varios cientos hasta más de mil quinientos dólares por ítem.

Las funciones de valoración de items comúnmente empleadas en los TAIs tienden a dificultar el mantenimiento del banco. Estas funciones de valoración suelen implicar un uso intensivo de aquellos items de mayor parámetro de discriminación. Si por cada ítem retirado (situado en la cola derecha en la distribución de discriminación) introducimos un nuevo ítem en el banco (cuya discriminación esperada será igual a la discriminación promedio del banco), lo que estaremos haciendo será ir reduciendo progresivamente la capacidad de discriminación de nuestro banco (Hau & Chang, 2001). Para evitar este deterioro, la ratio entre el número de items a calibrar para substituir a los eliminados y el número de items descartados tendrá que ser marcadamente mayor de 1. Por otro lado, una parte importante de las reglas de selección de items empleadas en los TAIs conlleva que gran parte del banco de items sea trivial, en el sentido de que nunca es administrado a examinado alguno. Podemos tener, pues, un gran coste para mantener la calidad del banco y un retorno nulo o mínimo de la inversión para el desarrollo de una parte importante del mismo. Esta situación puede disparar el coste de mantenimiento del programa de evaluación (Wainer, 2000b).

1.3.5. Importancia relativa de los diferentes objetivos

La relevancia de los cuatro objetivos descritos dependerá de multitud de factores específicos de cada programa de evaluación. Así, por ejemplo, fa fiabilidad deseada de las puntuaciones depende del uso que se le quiera dar a las mismas, al igual que la seguridad del banco de items es un asunto menor en aquellos tests en los que los examinados no tienen motivos para falsear sus respuestas.

En líneas generales, el primer objetivo, fiabilidad, y el segundo y el cuarto, seguridad y mantenimiento, son contrapuestos: incrementos en la satisfacción de uno suponen decrementos en los resultados en los otros. Una gran proporción de items de un banco no son óptimos para ningún nivel de rasgo en términos de información aportada. Por ello, cuando se prioriza la fiabilidad como criterio, una parte importante del banco no es usada

para ningún examinado. Este modo de proceder entra en conflicto con los objetivos de seguridad y mantenimiento. El control de contenidos tendrá un efecto menor en el resto de objetivos, asumiendo un banco de items bien construido en el que la composición del banco refleje las restricciones de contenido que se van a imponer.

Capítulo 2.
Estudios presentados.

2.1. Breve descripción de los estudios presentados

Esta tesis está compuesta por cuatro estudios, todos ellos centrados en la selección de items y seguridad en los TAIs. Los dos primeros corresponden a artículos publicados, uno en el 2008, el otro en el 2009. Los otros dos estudios se encuentran, en este momento, en proceso de revisión. Los dos primeros presentan vías para mejorar la seguridad de los TAIs. El primero es una extensión de la lógica de limitar la tasa máxima de exposición de los items. En el segundo se muestra cómo flexibilizar funciones de valoración presentadas anteriormente, de tal modo que se mejora la seguridad del banco mientras que se controla el impacto en el deterioro de la precisión. En el tercero, proponemos un método más riguroso para comparar diferentes reglas de selección de items, aceptando las variables dependientes comúnmente empleadas hasta el momento en la evaluación de TAIs. En el cuarto, se evalúa la validez de los indicadores de seguridad de los TAIs empleados tradicionalmente.

2.1.1. Múltiples tasas máximas de exposición en tests adaptativos informatizados

Hasta el momento, la regla de selección de items más comúnmente empleada, la de máxima información puntual, cuando se combina con los métodos que fijan un tasa máxima de exposición que ningún ítem ha de sobrepasar (Chen & Lei, 2005; Chen et al., 2008; Sympson & Hetter, 1985; Revuelta & Ponsoda, 1998; van der Linden & Veldkamp, 2004), supone que los items de mayor parámetro de discriminación son administrados al comienzo del test. Suponiendo que todos los examinados comenzaran el TAI con el mismo nivel de rasgo estimado, el ítem más informativo para ese nivel de rasgo sería presentado el "r^{max} por ciento" de las veces como primer ítem del test; igualmente, ese ítem nunca sería presentado en una posición que no fuera la primera. Sin embargo, diferentes estudios han mostrado que una selección de items altamente basada en el azar al comienzo del test apenas deteriora la estimación final de los niveles de rasgo (Li & Schafer, 2005; Revuelta & Ponsoda, 1998).

Nuestra propuesta es determinar tantos valores de r^{max} como items van a ser administrados. La tasa máxima de exposición de los items al comienzo del test será tan baja como resulte posible. La tasa máxima de exposición condicionada a la posición del ítem en el test irá incrementándose según avance el test. De este modo, esperamos mejorar la seguridad del banco sin reducir la fiabilidad del test.

2.1.2. Incorporando aleatoriedad a la información de Fisher para la mejora del control de la exposición en TAIs

Según el método progresivo (Revuelta & Ponsoda, 1998), la selección de items se realiza mediante la combinación de dos elementos, uno basado en la información de Fisher y otro en un componente aleatorio. El peso del componente aleatorio es máximo al comienzo del test y se va reduciendo linealmente según el test avanza. Revuelta y Ponsoda pusieron a prueba únicamente una función que relacionara la posición del ítem en el test y el peso del componente aleatorio. Sus resultados indican que mediante el método progresivo se mejora la seguridad del banco sin deteriorar la precisión. Así, queda la duda de si es posible dar un peso aún mayor a la aleatoridad en la selección manteniendo la fiabilidad. En este estudio, nos proponemos ensayar otras funciones para determinar relación entre posición del ítem en el test y el peso del azar en la selección. Específicamente, proponemos una que incorpora un parámetro que fija la velocidad con la que el componente aleatorio disminuye. De este modo, podemos evaluar qué niveles de aleatoriedad resultan tolerables en un TAI.

En el método proporcional (Segall, 2004), la información de Fisher determina la probabilidad de selección de los items. De este modo, items escasamente informativos para todos los niveles de rasgo tendrán tasas de exposición mayores de cero, lo cual no ocurre cuando se aplica máxima información puntual. Esta propuesta, por el momento, apenas ha recibido atención en la investigación sobre TAIs. Una posible extensión del método proporcional es emplear la información de Fisher elevada a una cierta potencia para determinar la probabilidad de selección. Si esta potencia es igual a 0, la selección es aleatoria. Cuanto mayor es la potencia, obtenemos resultados más parejos entre la selección mediante máxima información y el método proporcional. Nuestra propuesta es el empleo de una función que relacione la posición del ítem en el test y el valor del exponente, de tal modo que la aleatoriedad en la selección sea máxima al comienzo y vaya reduciéndose según avanza el test.

Por tanto, nuestro objetivo es el estudio de los niveles de azar que son tolerables en la selección de items, tanto con el método progresivo como con el proporcional, al mismo tiempo que ofrecemos flexibilizar los métodos, en comparación con sus propuestas originales.

2.1.3. Un método para la comparación de reglas de selección de ítems en tests adaptativos informatizados

Consideremos que únicamente son dos los objetivos de un TAI: una medida precisa del nivel de rasgo de los examinados (primer objetivo) mientras que se mantiene la seguridad del banco de items (segundo objetivo). Hasta la fecha, y como hemos visto, se han propuesto multitud de reglas de selección de items. Unas hacen mayor énfasis en el primer objetivo; otras, en el segundo.

Para que una regla pudiera considerarse como preferible a otra haría falta que, ofreciendo resultados iguales o mejores en un objetivo, obtuviera resultados mejores en el otro objetivo. En los estudios de comparación entre reglas de selección, este patrón de resultados no es común: las reglas mejores en seguridad acostumbran a ser peores en precisión. Tomemos como ejemplo unos datos de Chang y Ying (1999; Tabla 1, Estudio 2). En esa simulación, el método alfa-estratificado obtiene un RMSE (*root mean squared error*; raíz del error cuadrático medio) para la recuperación del nivel de rasgo igual a 0.29 y el método de máxima información un RMSE de 0.24. La tasa de solapamiento para el método estratificado es 0.10 y, en el caso de máxima información, la tasa es igual a 0.18. Ninguna de las dos reglas es mejor en ambos criterios a la vez. Este tipo de resultados acostumbran a comentarse con descripciones del estilo de "el incremento en el error de medida puede considerarse como poco importante en comparación con la mejora en la seguridad del banco".

Desde nuestro punto de vista, estos estudios ofrecen una información limitada a la hora de decidir qué regla es la que conviene aplicar en un TAI. Es conocido que cualquiera de los métodos de restricción de r^{max} permite reducir la tasa de solapamiento con el coste de incrementos en el RMSE. En el estudio de Chang y Ying (1999), habría sido posible probar diferentes valores de r^{max} hasta encontrar aquel con el que las dos reglas de selección se igualaran en una de las variables dependientes para, así, poder compararlas únicamente en la otra variable dependiente.

La complejidad del problema se incrementa cuando son más de dos las reglas de selección que deseamos comparar. En este caso, es probable que la búsqueda por

ensayo y error de valores que permitieran la comparación no sea practicable. Por ello, en este estudio proponemos una aproximación más sistemática al problema, similar a la ofrecida por Barrada et al. (2008) para comparar las opciones de bancos rotatorios y de restricción de tasa máxima. Nuestra propuesta es simular un amplio número de valores diferentes de r^{max} por cada regla de selección, de tal modo que se puede representar la relación entre r^{max} y tasa de solapamiento, por un lado, y entre r^{max} y RMSE, por otro. Con esta información, dibujar la gráfica que relaciona la tasa de solapamiento con el RMSE es sencillo. La representación simultánea de estos datos para diferentes reglas de selección permite saber, para cualquier valor posible de tasa de solapamiento o de RMSE, qué regla es la que ofrece un mejor resultado en la otra variable dependiente.

2.1.4. Filtrado del banco de items en tests adaptativos informatizados. ¿Qué hace más segura a una regla de selección de items?

Dos han sido los indicadores básicos de seguridad de un TAI: la tasa de solapamiento y la distribución de las tasas de exposición (Chang & Ying, 2002). Se ha asumido que bajas tasas de solapamiento y distribuciones más homogéneas de tasas de exposición suponen una mayor seguridad. Hasta el momento, esta idea no ha sido puesta en duda en la investigación relativa a los TAIs. Hay, sin embargo, motivos teóricos que permiten cuestionar la validez de estas variables en la evaluación de la seguridad. El principal problema se encuentra en que los estudios de evaluación de la seguridad se han realizado en condiciones donde los examinados carecen de información previa acerca del contenido de los items. Esto es, se ha evaluado seguridad del banco en condiciones donde no hay filtrado de items.

Tomemos como ejemplo el método estratificado y supongamos: (a) que todos los examinados comienzan con el mismo nivel de rasgo estimado; y (b) que no se aplica restricción en la tasa máxima. Según estudios previos, este modo de proceder implicaría una tasa de solapamiento bastante próxima a la minima posible. Con esta regla de selección, examinados con niveles de rasgo alejados compartirán un número reducido de items y los pocos items compartidos se concentrarán en las primeras posiciones del test. Imaginemos ahora que un examinado con un nivel de rasgo bajo obtiene información sobre los items que recibió un examinado de rasgo alto. Puesto que, por diseño, ambos examinados compartirán el primer ítem, el de rasgo bajo acertará un ítem que podíamos esperar que hubiera fallado. El modelo de TRI con el que se calibraron los items ya no describe convenientemente las probabilidades de respuesta correcta. El examinado de

bajo nivel de rasgo compartirá también el segundo ítem con el examinado de alto nivel de rasgo. Es de esperar que ambos acierten este ítem. Lo que, según los estudios tradicionales, supondría una baja tasa de solapamiento, se convierte en un alto número de items compartidos y en un gran beneficio por afrontar el examen con información previa.

Los métodos progresivo y proporcional ofrecen tasas de solapamiento superiores a las obtenidas con el método estratificado, si bien el solapamiento es mínimo al comienzo del test, puesto que se comienza con selección aleatoria. La probabilidad de que un examinado de bajo nivel de rasgo pueda beneficiarse de la información recibida por un examinado de alto nivel de rasgo es baja: la probabilidad de compartir los primeros items es mínima.

Con este ejemplo, se ilustra la necesidad de evaluar los TAIs en condiciones de difusión del contenido de los bancos de items. De este modo, la simulación se aproxima en mayor medida a lo que en la realidad puede suceder cuando se quiebra la seguridad del banco de items. Con este objetivo en mente, se ofrecen diferentes estudios. En el primero de ellos, se realiza una comparación entre reglas de selección según el modo aplicado hasta el momento. En los siguientes, se muestra cómo las reglas que, según el primer estudio, habían de ser las más resistentes a la difusión del contenido del banco, no lo son: las reglas de menor tasa de solapamiento no son las que menor distorsión sufren cuando los examinados se presentan al examen con información previa acerca del contenido del banco.

2.2. Relación entre estudios

En el primer y segundo estudio presentamos desarrollos y mejoras orientadas a mejorar la seguridad del banco. Una de ellas, referida al modo de construir B_q mediante el control de r^{max}. La otra, mediante dos propuestas de función de valoración de items. En estos dos primeros estudios, damos por correcto el modo tradicional de proceder sobre reglas de selección, aceptando las variables dependientes habitualmente usadas y el balance entre precisión y seguridad. En el tercer estudio, cuestionamos el alcance de las conclusiones que pueden extraerse de investigaciones en las que se encuentra tal balance. Este problema afectaría también a nuestros dos primeros estudios. Para solucionar esta limitación proponemos un método más exhaustivo mediante el cual se puede comparar diferentes métodos garantizando que únicamente difieren en una variable. En este tercer

estudio, seguimos dando por buenas las variables dependientes que miden precisión y seguridad, si bien damos un primer paso en la mejora de la comparación entre reglas. En el cuarto estudio, nos planteamos la validez de la tasa de solapamiento y la distribución de las tasas de exposición como medidas de seguridad del banco. Estas variables son, por lo general, evaluadas en condiciones en las que se asume que el banco es totalmente seguro. Argumentamos que, si se desea evaluar seguridad, lo correcto es simular un entorno de filtrado de items, señalando motivos por los que las variables usuales pueden llevar a conclusiones incorrectas. Este es un nuevo paso en la comparación de reglas de selección de items. En general, podríamos apuntar a que los estudios presentados van desde la mejora de lo comúnmente aplicado y aceptado en TAIs hasta la reevaluación y redefinición del modo de investigar en este campo.

2.3. Opciones de investigación

Hay TAIs de longitud fija y de longitud variable; hay TAIs de escalamiento y de clasificación; los hay con modelos dicotómicos de TRI, politómicos, con *testlets* (conjuntos de preguntas con un enunciado común), no paramétricos o multidimensionales... La mayor parte de la investigación publicada toma los parámetros de los items como carentes de error de estimación, mientras que otros estudios evalúan el efecto de estos errores (Li & Schafer, 2003; van der Linden & Glas, 2000).

En este amplio panorama de opciones, era necesario tomar algunas decisiones a la hora de llevar adelante los estudios que ahora se presentan. El criterio fundamental fue empezar por un modelo sencillo y de amplia aplicación, como es el modelo logístico de 3 parámetros. Hoy por hoy, la mayor parte de los trabajos están basados en tests de longitud fija con interpretación referida a norma. La investigación en TAIs con estas premisas es extensa, por lo que las comparaciones de los resultados nuevos con lo esperado según artículos previos se facilita. Adicionalmente, la extensión de lo presentado a otros modelos de TRI o demás configuraciones de un TAI resulta relativamente directa y sencilla.

En tres de los cuatro estudios que presentamos no hay restricción de contenidos, pese a haberlo indicado como uno de los cuatro objetivos básicos de un TAI. Esto se debe a que para un banco bien construido, los efectos de la restricción son mínimos con respecto a seguridad y precisión. Por tanto, se espera que los resultados encontrados sin control de contenidos sean extrapolables a condiciones con restricciones. Adicionalmente, nuestro

punto de vista es que resulta más oportuno comenzar investigando condiciones sin restricción para, posteriormente, ir introduciendo otros elementos.

La facilidad de mantenimiento del banco de items, también enunciado como un objetivo general, ha sido evaluada directamente en sólo uno de los estudios. Por el momento, este objetivo ha recibido una atención menor en los estudios sobre TAIs y se asume que mejoras en la seguridad del banco facilitan el reemplazo de items.

Una parte importante de la literatura sobre TAIs se realiza empleando bancos de items reales, cuyos parámetros resultan inaccesibles. Esto plantea problemas a la hora de intentar replicar de resultados por otros grupos de investigación y deja la duda de si ciertos discrepancias entre estudios pueden deberse a diferencias en la composición de los bancos de preguntas. Así, por ejemplo, Barrada, Olea et al. (2009) han mostrado que la correlación entre el parámetro de discriminación y el parámetro de localización de los items puede suponer cambios en el funciomiento relativo de diferentes reglas de selección (reglas que resultan recomendables en ausencia de correlación pasan a ser rechazables en presencia de ésta). Por esto, los cuatro estudios presentados incluyen simulaciones realizadas con bancos de items cuyos parámetros han sido generados aleatoriamente y se informa de las distribuciones generadoras. En dos de los estudios, se evalúa si los resultados se generalizan al tomar como parámetros de los items los parámetros de un banco de items actualmente operativo, eCAT (Olea et al., 2004).

PARTE II – ESTUDIOS

Capítulo 3.
Múltiples tasas máximas de exposición en tests adaptativos informatizados

Artículo publicado. La referencia del mismo es:

Barrada, J. R., Veldkamp, B. P., & Olea, J. (2009). Multiple maximum exposure rates in computerized adaptive testing. *Applied Psychological Measurement, 33*, 58-73.

Multiple Maximum Exposure Rates in Computerized Adaptive Testing

Juan Ramón Barrada, Universidad Autonoma de Barcelona
Bernard P. Veldkamp, University of Twente
Julio Olea, Universidad Autonoma de Madrid

Computerized adaptive testing is subject to security problems, as the item bank content remains operative over long periods and administration time is flexible for examinees. Spreading the content of a part of the item bank could lead to an overestimation of the examinees' trait level. The most common way of reducing this risk is to impose a maximum exposure rate (r^{max}) that no item should exceed. Several methods have been proposed with this aim. All of these methods establish a single value of r^{max} throughout the test. This study presents a new method, the multiple-r^{max} method, that defines as many values of r^{max} as the number of items presented in the test. In this way, it is possible to impose a high degree of randomness in item selection at the beginning of the test, leaving the administration of items with the best psychometric properties to the moment when the trait level estimation is most accurate. The implementation of the multiple-r^{max} method is described and is tested in simulated item banks and in an operative bank. Compared with a single maximum exposure method, the new method has a more balanced usage of the item bank and delays the possible distortion of trait estimation due to security problems, with either no or only slight decrements of measurement accuracy.
Index terms: computerized adaptive testing, item exposure control, test security, item selection

Computerized adaptive testing (CAT) of knowledge, abilities, and skills offers several advantages. CATs are administered individually and they are flexible. Moreover, they are more efficient than traditional paper-and-pencil testing in that the difficulty of the items can be adapted to the proficiency of the candidate (Segall, 2004).

However, CATs have also been criticized. First of all, they are subject to security problems. When they are online, they are vulnerable to (organized) item theft. Candidates might memorize items and publish them on the Internet or simply share them with friends who might take the CAT in the future (Chang, 2004). A second problem is related to item bank usage. Many items in the banks are rarely selected for administration, because most item selection rules favor other items for their better measurement qualities. Thus, both time and money are wastefully invested in developing them.

Both problems can be formulated in terms of exposure of individual items: security problems are related to variance in the exposure rates of the items (Chen, Ankenmann, & Spray, 2003); poor item bank usage is related to an underexposure of less popular items. To deal with these problems, various exposure control methods have been proposed, the most popular being that of Sympson and Hetter (1985). Numerous modifications of this method have been presented (Stocking & Lewis, 1998; van der Linden, 2003). Chang and Ying (1999) proposed the alpha-stratified method;

Applied Psychological Measurement, Vol. 33 No. 1, January 2009, 58–73
DOI: 10.1177/0146621608315329

Revuelta and Ponsoda (1998) the progressive method, which focuses on underexposure problems; and more recently, van der Linden and Veldkamp (2004, 2007) developed the item-eligibility method.

In all these methods, the following trade-off can be found: the greater the emphasis on exposure control, the greater are the costs in terms of measurement precision (Way, 1998). From the inverse point of view, the more accurate the CAT, the higher are the risks to the item bank. In fact, this study deals with a multiple-criteria decision-making problem. The first criterion is measurement precision; the second, exposure control. Therefore, the challenge in developing or selecting exposure control methods lies in finding the method that performs best with respect to both measurement precision and observed exposure rates or test overlap.

In this article, a new method for exposure control, the multiple-$r^{\max}$ method (MRM), is described. In this method, exposure control parameters are varied throughout the test administration. It is argued that increased item bank usage can be achieved with this method, with either no or at the most only minimal increments in measurement errors.

The common method for improving bank security is to control the maximum exposure rate. First, one of the methods for doing so, the item-eligibility method (van der Linden & Veldkamp, 2004), is described. After that, some of its limitations are shown and then the rationale of the MRM method and its implementation are presented. Two simulation studies are described, one with randomly generated item banks and the other with an operative item bank.

Item-Eligibility Method

The goal of methods that control the *maximum exposure rate* is to set all *item exposure rates* below a maximum exposure rate, $r^{\max}$, fixed beforehand by the testing agency:

$$P(A_i) \leq r^{\max}, \tag{1}$$

where $Q/n \leq r^{\max} \leq 1$ (Chen et al., 2003), $P(A_i)$ is the probability of administering the ith item, Q is the number of items to be administered, and n is the item bank size. All the methods introduce exposure control parameters for the items. The first method presented was the Sympson–Hetter method (Sympson & Hetter, 1985). This approach involves a time-consuming process to fix the exposure control parameters (Barrada, Olea, & Ponsoda, 2007; Chen & Doong, 2008; van der Linden, 2003). Some methods have been proposed that adapt the control parameters on the fly. The restricted method (Revuelta & Ponsoda, 1998) has this characteristic, but some drawbacks of the method have been described (Chen, Lei, & Liao, in press). Recently, van der Linden and Veldkamp (2004, 2007) described the item-eligibility method, which the present study uses as a benchmark for comparison with the MRM method.

In the item-eligibility method, two events are defined: (a) item i is eligible for the examinee (E_i) and (b) item i is administered (A_i). Exposure control is achieved by restricting the proportion of examinees for which an item can be eligible. This proportion is $P(E_i)$.

The $P(E_i)$ values are adapted on the fly for each new examinee. The parameters for the $(j+1)$-th examinee—$P(E_i^{j+1})$—are calculated using the following equation:

$$P(E_i^{j+1}) = \begin{cases} 1 & \text{if} \quad P\left(A_i^{1..j}\right)/P\left(E_i^{j}\right) \leq r^{\max} \\ r^{\max} P\left(E_i^{j}\right)/P\left(A_i^{1..j}\right) & \text{if} \quad P\left(A_i^{1..j}\right)/P\left(E_i^{j}\right) > r^{\max} \end{cases}, \tag{2}$$

where $P(A_i^{1..j})$ is the exposure rate (probability of administration) of the ith item in the range of examinees between the first and the jth examinee.

For each examinee, a subset of eligible items is formed before any item is administered. For each item, a random number belonging to the uniform interval (0, 1) is generated. Only if that number is smaller than $P(E_i^{j+1})$ is that item eligible. During the administration of the test, only eligible items can be administered. This way, except for a few random exceptions, all items have exposure rates equal to or below $r^{\max}$.

The methods presented to date for controlling the maximum exposure rate share several drawbacks. Assuming that the item with maximum Fisher information for the estimated trait level is selected and items calibrated according to the three-parameter logistic model are used, items with a high a parameter and a low c parameter from the beginning of the test will be chosen (Barrada, Olea, Ponsoda, & Abad, 2006) when the estimation of the trait level is unstable and measurement error is high. As items in the bank with this combination of parameters are infrequent, the quality of items measured by the information they provide will decline as the test goes on (Revuelta & Ponsoda, 1998).

Even with $r^{\max}$ values close to the minimum possible, the mean of the a parameter of the items administered is still above the mean of the a parameter in the bank (Barrada et al., 2007). The methods of restriction of the maximum exposure rate reduce the exposure rate of overexposed items while increasing the exposure rate of items whose exposure rates are smaller and closer to $r^{\max}$.

Hau and Chang (2001) have shown that it is advisable to increase the value of the a parameter of the items administered as the test goes on, instead of reducing it. In fact, the random selection of items at the beginning of the test and subsequent selection based on Fisher information means no reduction or a very small reduction (Barrada, Olea, Ponsoda, & Abad, in press; Li & Schafer, 2005; Revuelta & Ponsoda, 1998) in measurement accuracy.

One method for balancing the item exposure rates would involve establishing as many maximum exposure rates as items to be administered (Q). A proposal for doing so is presented in the next section.

A New Method: The Multiple-$r^{\max}$ Method

The goal of the multiple-$r^{\max}$ method (MRM) can be seen in equation (3):

$$P(A_{i,1..q}) \leq r_{1..q}^{\max}, \tag{3}$$

where $r_{1..q}^{\max}$ is the desired maximum exposure rate until the qth item and $P(A_{i,1..q})$ is the exposure rate of the i item considering the first q items in the test.

The definition of the $r_{1..q}^{\max}$ values is subject to the following restrictions:

$$r_{1..q+1}^{\max} > r_{1..q}^{\max}, \tag{4}$$

$$r_{1..q}^{\max} \geq q/n, \tag{5}$$

and

$$r_{1..Q}^{\max} \leq 1. \tag{6}$$

If $r_{1..q+1}^{\max}$ was allowed to be equal to $r_{1..q}^{\max}$, those items with $P(A_{i,1..q})$ equal to $r_{1..q}^{\max}$ could not be administered in the $(q+1)$-th position of the test. This is avoided by the first restriction. The other two restrictions mark the limits between which the $r_{1..q}^{\max}$ values have to be.

The lowest $r_{1..q}^{\max}$ is imposed at the beginning of the test. In this way, it is possible to avoid selecting all items with high a and low c parameters when estimation of trait levels is still unstable. The

values of $r^{\max}_{1..q}$ increase during CAT administration, which implies that more items become eligible.

The MRM model presented by inequalities 3 to 6 is not a stand-alone method of item exposure control but a general structure that needs to be combined with a method for controlling maximum exposure rates. The next section explains how the MRM has been implemented, combined with the item-eligibility method (van der Linden & Veldkamp, 2004).

Implementation of MRM

Some examples are first presented to illustrate the implementation of this new method. Imagine the following condition: $r^{\max}_{1..q-1}$ equals 0.24 and $r^{\max}_{1..q}$ equals 0.25. Which maximum exposure rate can therefore be tolerated for the qth position? Not 0.25. If item 1 is exposed to 24% of the examinees until the $(q-1)$-th position and exposed to 25% in the next position, the total exposure rate of that item would clearly be over our limit.

The two maximum exposure rates for an item position need to be differentiated. First, the maximum exposure rate acceptable during the first qth items is considered. As stated above, this first maximum exposure rate is termed $r^{\max}_{1..q}$. Second, the maximum exposure rate tolerable in the qth position.

This tolerable exposure rate could be defined as the difference between $r^{\max}_{1..q}$ and $r^{\max}_{1..q-1}$. In this case, in the present example, the tolerable rate for the qth item position would be 0.01. Consider item 2, exposed to 5% of the examinees until the $(q-1)$-th position, not reaching $r^{\max}_{1..q}$. If an exposure rate of only 0.01 is tolerated in the qth position, the total exposure rate of this item would be markedly below the limit, losing tolerable usage of that item. This article attempts to satisfy equation (3) without overrestricting exposure rates. It can be seen with the example of item 2 that the tolerable rate should be dependent on the exposure rate in the previous item position.

One option might be to calculate the tolerable exposure rate for the qth position equal to $r^{\max}_{1..q}$ minus the actual exposure rate until the $(q-1)$-th position. For item 1, this value is equal to 0.01 and for item 2 it is equal to 0.2. For these two items there is no problem with this definition of the tolerable exposure rate for the qth position. But imagine items 3 and 4. Item 3 is exposed to 0.245 of the examinees until the $(q-1)$-position and item 4 to 0.26. Both exposure rates are greater than $r^{\max}_{1..q-1}$ and in the case of item 4, greater than $r^{\max}_{1..q}$. Observed exposure rates higher than those desired are possible because the process includes a random component. If the definition of tolerable exposure rate as $r^{\max}_{1..q}$ minus the actual exposure rate until the $(q-1)$-th position is applied, the figures for items 3 and 4 would be 0.005 and -0.01. It is meaningless to set a negative value for the tolerable exposure rate in an item position. One option could be to fix negative values at zero, but the convenience of not doing so has been defended in this article (equation (4)). Both the MRM method and the item-eligibility method adapt all the parameters on the fly to try to satisfy the restrictions imposed, so it makes sense to suppose that when a new examinee is tested, the observed exposure rates of the items that exceed exposure limits will fall to the limits fixed. Considering that this control is achieved, the tolerable item exposure rate would then be 0.01 for items 3 and 4. In this way, it can be seen how the tolerable exposure rate will depend on the estimation of the exposure rate in the previous item position when a new examinee is tested. This estimation is made as in equation (7):

$$\hat{P}\left(A^{1..j+1}_{i,1..q}\right) = \begin{cases} P\left(A^{1..j}_{i,1..q}\right) & \text{if} \quad P\left(A^{1..j}_{i,1..q}\right) < r^{\max}_{1..q} \\ r^{\max}_{1..q} & \text{if} \quad p\left(A^{1..j}_{i,1..q}\right) \geq r^{\max}_{1..q} \end{cases} \tag{7}$$

Table 1
Examples of Definition of Maximum Exposure Rate in a Position of the Test

	a		*b*		*c*		*d*	
q	1	2	1	2	1	2	1	2
$r^{\max}_{1..q}$	0.05	0.15	0.05	0.15	0.05	0.15	0.05	0.15
$P(A^{1..j}_{i,1..q})$	0.03	0.15	0.05	0.20	0.10	0.15	0.18	0.20
$t^{j+1}_{i,q}$	0.05	0.12	0.05	0.10	0.05	0.10	0.05	0.10

As can be seen, it is assumed that in the event the exposure rate for one examinee exceeds the maximum exposure rate, the exposure control method will be able to restrict the exposure rate to $r^{\max}_{1..q}$ for the next examinee.

The tolerable rate in the qth position in the test for the ith item in the bank for the $(j+1)$-th examinee is referred to as $t^{j+1}_{i,q}$. The value of $t^{j+1}_{i,q}$ will be $r^{\max}_{1..q}$ minus the estimation of $P(A^{1..j+1}_{i,1..q})$. Thus, the equation for calculating $t^{j+1}_{i,q}$ is as follows:

$$t^{j+1}_{i,q} = r^{\max}_{1..q} - \hat{P}\left(A^{1..j+1}_{i,1..q-1}\right) = r^{\max}_{1..q} - \min\left[r^{\max}_{1..q-1}, P\left(A^{1..j}_{i,1..q-1}\right)\right]. \tag{8}$$

Table 1 shows four examples presenting how $t^{j+1}_{i,q}$ is calculated. In example *d*, it is clear why it is better to use equation (7) for calculating $\hat{P}(A^{1..j+1}_{i,1..q})$ instead of simply making $\hat{P}(A^{1..j+1}_{i,1..q})$ equal to $P(A^{1..j}_{i,1..q})$. With the option chosen, it is impossible for $t^{j+1}_{i,q}$ to be negative or equal to zero.

When MRM is combined with the item-eligibility method, the control parameters for item i for the $(j+1)$-th examinee and the qth item position is calculated according to the following equation:

$$P\left(E^{j+1}_{i,q}\right) = \begin{cases} 1 & \text{if} \quad P\left(A^{1..j}_{i,1..q}\right)/P\left(E^{j}_{i,q}\right) \le t^{j+1}_{i,q} \\ P\left(E^{j}_{i,q}\right) t^{j+1}_{i,q} / P\left(A^{1..j}_{i,1..q}\right) & \text{if} \quad P\left(A^{1..j}_{i,1..q}\right)/P\left(E^{j}_{i,q}\right) > t^{j+1}_{i,q} \end{cases} \tag{9}$$

As can be seen, this equation is similar to the one used for calculating the $P(E_i)$ parameters in the item-eligibility method but $r^{\max}$ is replaced with $t^{j+1}_{i,q}$.

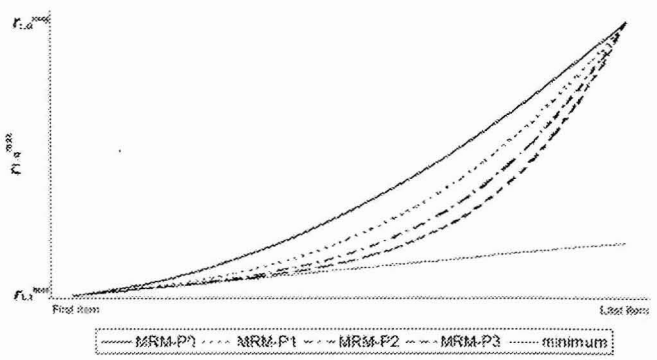
Once the control parameters are calculated, it is possible to define which part of the bank is eligible for each item position. This is done as explained in the item-eligibility method, with the difference that eligibility is not defined for the test but for each item position. For doing so, n random numbers in the uniform interval (0, 1) are generated, one for each item position, and if these numbers are lower than the control parameters the items are eligible for those item positions.

If content constraints are incorporated into the test (van der Linden, 2005), there is a possibility of no feasible test existing. This would happen when the eligible items cannot meet the content specifications. Van der Linden and Veldkamp (2004) discuss how to incorporate the probability of infeasibility into the item-eligibility method. For an item bank correctly constructed, this probability is considered to be very small and this element is not introduced in the present method.

A Possible Function for Defining the Values of $r^{\max}_{1..q}$

The random selection of items at the beginning of the test has a small impact on measurement accuracy (Barrada et al., in press; Li & Schafer, 2005; Revuelta & Ponsoda, 1998). Thus, it seems

Figure 1
Examples of Functions Relating Item Position to $r_{1..q}^{max}$ for Four Acceleration Parameters and the Minimum Admissible Values of $r_{1..q}^{max}$



appropriate to strongly adjust the $r_{1..q}^{max}$ values for the first items to the minimum admissible values (equation (5)), as this would improve the balanced usage of the item bank. The value for $r_{1..Q}^{max}$ would be set to the value that for security reasons the testing agency considers suitable. The most logical option is to fix $r_{1..Q}^{max}$ equal to the value that r^{max} would have if a method with a single r^{max} was applied.

A possible function for defining the $r_{1..q}^{max}$ values that makes $r_{1..1}^{max}$ equal to n^{-1} and leaves the freedom to set $r_{1..Q}^{max}$ is as follows:

$$r_{1..q}^{max} = \begin{cases} \frac{q}{n} & \text{if} \quad q = 1 \\ \left[1 + \frac{\left(r_{1..Q}^{max} \Big/ Q/n^{-1}\right) \sum_{h=2}^{q} (h-1)^s}{\sum_{h=2}^{Q} (h-1)^s}\right] \frac{q}{n} & \text{if} \quad q \neq 1 \end{cases}, \tag{10}$$

where h is a dummy variable only used for calculations and s is the acceleration parameter defining the speed with which $r_{1..q}^{max}$ separates from the minimum possible values for approaching $r_{1..Q}^{max}$. Examples of this function are shown in Figure 1.

With an acceleration parameter equal to zero, the ratio between $r_{1..q}^{max}$ and q/n (the minimum admissible value; equation (5)) increases in a linear fashion from 1 to $r_{1..Q}^{max} \Big/ Q/n$. The higher the value of the s parameter, the lower the speed with which the $r_{1..q}^{max}$ values increase.

It is important to note that obtaining a homogeneous distribution of the exposure rates by adjusting the values of $r_{1..q}^{max}$ to their minimum admissible values is not equivalent to selecting items randomly. By setting $r_{1..q}^{max}$ equal to q/n, the exposure rates for the overall population are homogenized, although when considering the exposure rates conditional on trait levels there will be a variance in the distribution of exposure rates that would not occur with random selection. Owing to this, the measurement error achieved with methods of restriction of maximum exposure rate—even though the maximum exposure rates are fixed at the minimum possible—will be lower than that found with random selection.

Two simulation studies were carried out to evaluate the performance of MRM as compared with a method using a single value of r^{max}, the item-eligibility method. In the first study, randomly generated item banks were used, whereas in the second a currently operative item bank for the assessment of knowledge of English grammar (Olea, Abad, Ponsoda, & Ximénez, 2004) was used.

Volume 33 Number 1 January 2009

Simulation Study 1

Method

Item banks and test length. Ten item banks of 500 items were generated. The distributions for the parameters were as follows: for a, $N(1.2, 0.25)$; for b, $N(0, 1)$; and for c, $N(0.25, 0.02)$. The test length was fixed at 25 items.

Trait level of the examinees. This study aimed to obtain the results for the overall population and conditional on several θ values. It was decided to sample nine θ values, ranging from -2 to 2 in steps of 0.5. To do so, a pool of examinees was constructed with the following two conditions: (a) the number of examinees in each θ value had to be proportional to the density at that point assuming a distribution $N(0, 1)$; (b) the minimum number of examinees for any θ value had to be equal to 1,000, considering this number large enough to obtain stable results. In this way, the pool of examinees was composed of 36,193 elements: for θ equal to -2 and 2, 1,000 examinees were sampled; for θ equal to -1.5 and 1.5, 2,399 examinees; for θ equal to -1 and 1, 4,482 examinees; for θ equal to -0.5 and 0.5, 6,521 examinees; and for θ equal to 0, 7,389 examinees. The trait level of each examinee simulated was randomly extracted without replacement from this pool.

Estimation of trait level and item selection rule. The starting $\hat{\theta}$ was fixed at 0. The estimator of θ was the expected a posteriori (EAP; Bock & Mislevy, 1982) estimator with a uniform prior over $[-4, 4]$. The selection algorithm most widely used in CATs was used: selecting the item with maximum Fisher information at the current estimated trait level.

$r_{1..q}^{\max}$ *values.* These values were adjusted for each item position in the test as shown in equation (10). The $r_{1..Q}^{\max}$ value was set equal to 0.25. Four values were used for the acceleration parameter: 0, 1, 2, and 3. In the item-eligibility method, $r^{\max}$ was equal to 0.25.

Performance measures. Six variables were used for the comparison between methods: (a) observed maximum exposure rates; (b) exposure rates of the items at the end of the test; (c) overlap rate, as defined in equation (11) (Chen et al., 2003); (d) RMSE (root mean square error), as shown in equation (12); (e) bias, calculated following equation (13); and (f) the information provided by the items for the real trait level of the examinee. RMSE and bias were calculated both for the whole set of simulees and conditional on the different θ values.

The overlap rate was

$$\hat{T} = \frac{n}{Q} S_{P(A)}^2 + \frac{Q}{n}, \tag{11}$$

where $\hat{T}$ is the large-sample approximation of the overlap rate (Chen et al., 2003) and $S_{P(A)}^2$ is the variance of the item exposure rates.

The RMSE and bias were

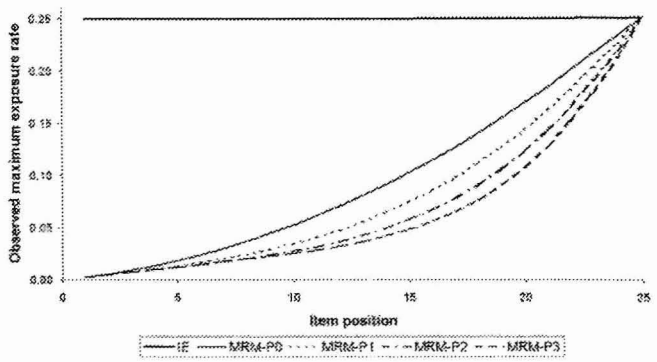
$$\text{RMSE} = \left(\sum_{g=1}^{m} \left(\hat{\theta}_g - \theta_g \right)^2 \Big/ m \right)^{1/2} \tag{12}$$

and

$$\text{Bias} = \sum_{g=1}^{m} \left(\hat{\theta}_g - \theta_g \right) \Big/ m, \tag{13}$$

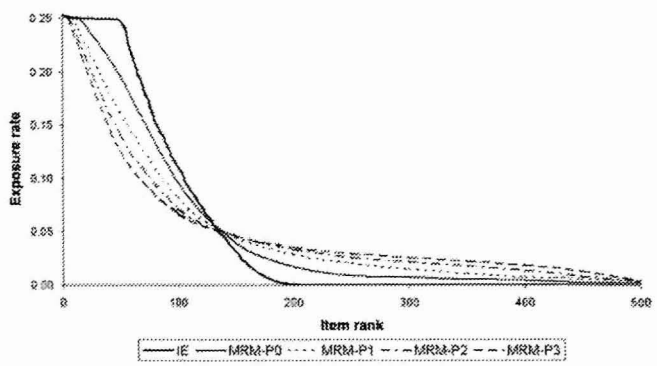
where m is the number of examinees, $\hat{\theta}_g$ is the estimated trait level for the gth examinee, and θ_g is the real trait level.

Figure 2
Observed Maximum Exposure Rate According to Item Position for the Theoretical Item Banks



Note: IE = item-eligibility method.

Figure 3
Exposure Rates of Items for Theoretical Item Banks



Note: The items are ordered according to their exposure rates. IE = item-eligibility method.

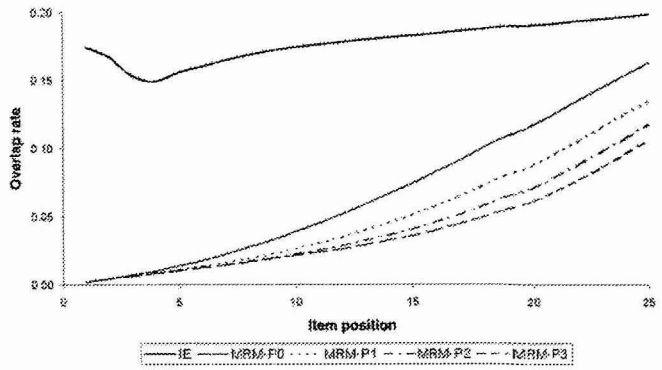
Results

Figure 2 shows the maximum exposure rates according to the item position in the test. With the item-eligibility method, the maximum exposure rate is already r^{max} for the first item and remains constant through the test. The MRM method shows the desired pattern for the different acceleration parameters studied: maximum rate at the beginning of the test is very low and increases as the number of items administered increases. The magnitude of this increase is controlled by the acceleration parameter. Both the item-eligibility method and the MRM method succeed in controlling the maximum exposure rate at the desired level.

As expected, the MRM method leads to a more homogeneous distribution of exposure rates, as can be seen in Figure 3. Although with the item-eligibility method about 60% of the items in the bank are never used, with the MRM method no item has an exposure rate equal to zero. MRM also

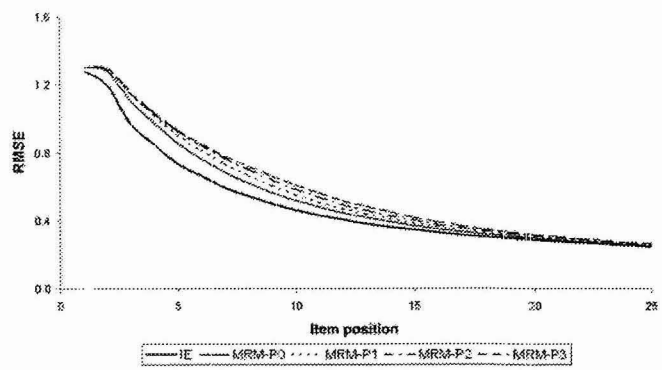
Volume 33 Number 1 January 2009

Figure 4
Overlap Rate According to Item Position for the Theoretical Item Banks



Note: IE = item-eligibility method.

Figure 5
RMSE According to Item Position for the Theoretical Item Banks



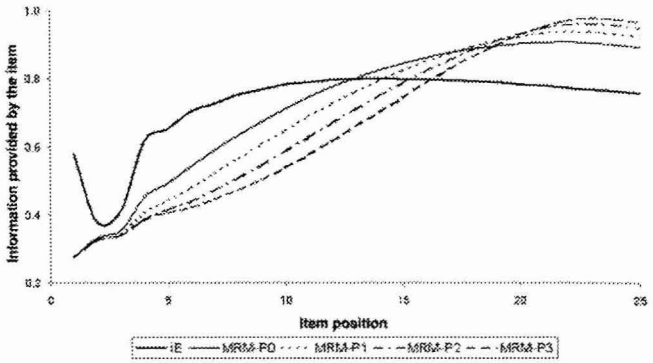
Note: IE = item-eligibility method; RMSE = root mean square error.

reduces the proportion of items with rates close to the maximum limit established. The greater the value of the acceleration parameter, the greater are the improvements in relation to under- and overexposure.

Greater exposure control means a reduction in the overlap rate achieved with the MRM method, which is always lower than the overlap rate obtained with the item-eligibility method, as can be seen in Figure 4. The overlap correlates negatively with the acceleration parameter. The differences between the item-eligibility method and MRM are greater at the beginning of the test and decrease as more items are administered.

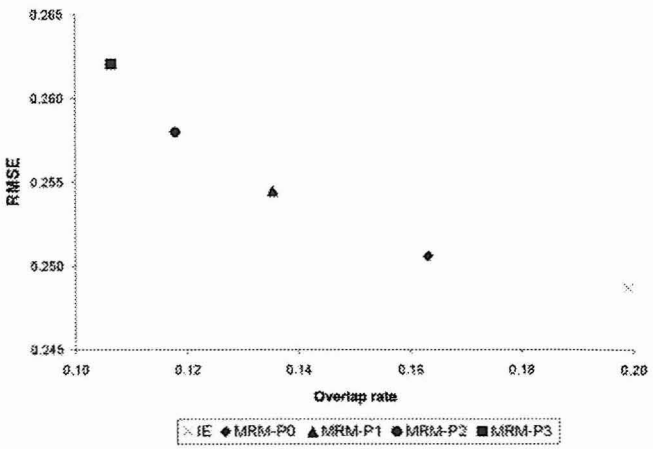
The lower exposure rate of MRM at the beginning of the test can be an additional advantage of the method. The distortion in the estimated trait level will be greater if the items that the examinee has previous knowledge of are at the beginning of the test, and especially if their *a* parameter is high (Chang & Ying, in press), as occurs with the item-eligibility method.

Figure 6
Fisher Information Provided by the Item According to Item Position for the Theoretical Item Banks



Note: IE = item-eligibility method.

Figure 7
Overlap Rate and RMSE for the Theoretical Item Banks



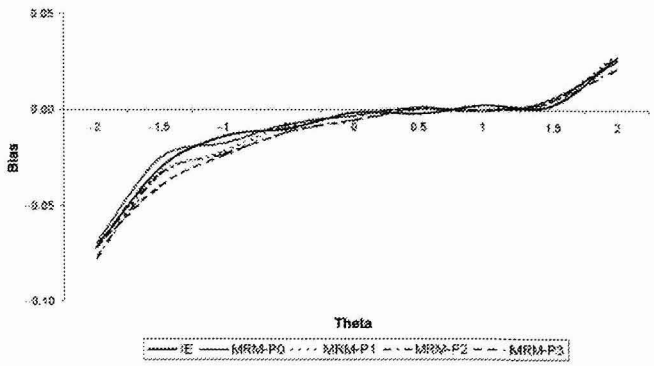
Note: IE = item-eligibility method; RMSE = root mean square error.

The effect of this greater exposure control on accuracy can be seen in Figure 5. At the beginning of the test, the MRM method offers a higher RMSE than the item-eligibility method, but this difference between them quickly falls, and by the end of the test it is almost unnoticeable.

The reason why it is possible to improve the bank security with almost no impact on accuracy can be seen in Figure 6, which shows the plotting of the mean information provided by each item for the examinee's real trait level. In the item-eligibility method, the information provided by each item increases in the first half of the test as the estimation approaches the real trait level. The exception of the second item in the test, when the information provided is reduced, is because the

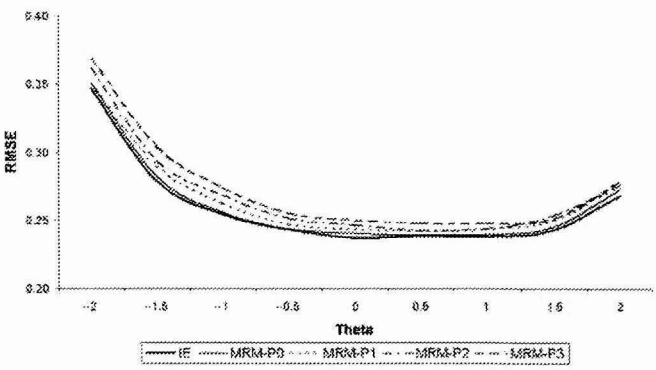
Volume 33 Number 1 January 2009

Figure 8
Bias According to θ Values for the Theoretical Item Banks



Note: IE = item-eligibility method.

Figure 9
RMSE According to θ Values for the Theoretical Item Banks

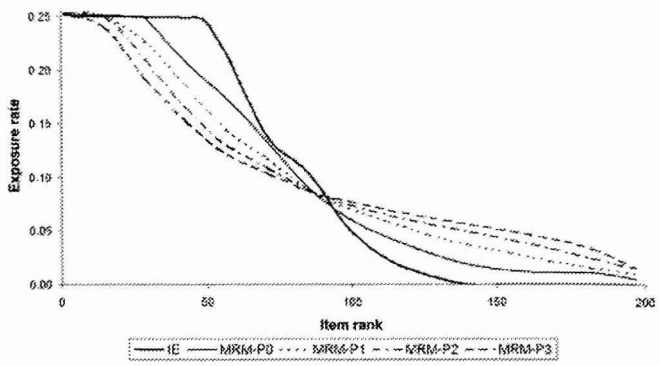


Note: IE = item-eligibility method; RMSE = root mean square error.

estimation after administration of just one item is necessarily far removed from the mean of the trait level in the population. For the second half of the test, the information provided by each item in the item-eligibility method reduces with each new item presented, as the trait level estimated is more stable and the highly informative items have already been used. Considering the first items of the test, the items presented with the MRM method are less informative than the items administered with the item-eligibility method. However, as far as the latter part of the test is concerned, there are still high-quality items available in the bank. This means that the information gathered with the two methods is similar and explains the small difference in RMSE.

Because the relevant point in practical settings is what is obtained at the end of the test, Figure 7 shows the overlap rates and RMSE for the two methods after 25 items. With regard to overlap, it shows how the MRM method clearly improves item bank security, compared with the item-eligibility method. This improvement increases as the acceleration parameter increases. With regard

Figure 10
Exposure Rates of Items for the Operative Bank



Note: The items are ordered according to their exposure rates. IE = item-eligibility method.

to RMSE it is clear how, as noted above, the differences are small and never greater than 0.015. The greater the value of the acceleration parameter, the greater is the RMSE.

Figure 8 shows the bias for the different methods according to the θ values sampled. The main differences are located for low trait-level values and can be considered as small. The higher the acceleration parameter, the higher is the bias. For θ values greater than 0, no differences are distinguishable. The same can be said for differences between the item-eligibility method and the MRM method with an acceleration parameter equal to 0.

The RMSE for the item-eligibility method is lower than the RMSE for the MRM method when acceleration parameters more than 0 are considered, as can be seen in Figure 9. When MRM with an acceleration parameter of 0 is compared with the item-eligibility method, the differences are negligible. The RMSE correlates positively with the acceleration parameter.

In short, the MRM method, compared with the item-eligibility method, allows for greater exposure control. It reduces the overlap rate and the number of items with exposure rates close to the limit rate and increases the exposure rate of the underexposed items. Moreover, these advantages are achieved with very little impact on measurement accuracy. The explanation is that the MRM method saves the most informative items for the latter part of the test, when the trait estimation is more accurate.

It seems clear that in the case of randomly generated item banks, the MRM method is an option that improves the security control of the item bank when compared with the methods that work with just one maximum exposure rate. To test whether the results of this study can be generalized, MRM was applied to a currently operative item bank.

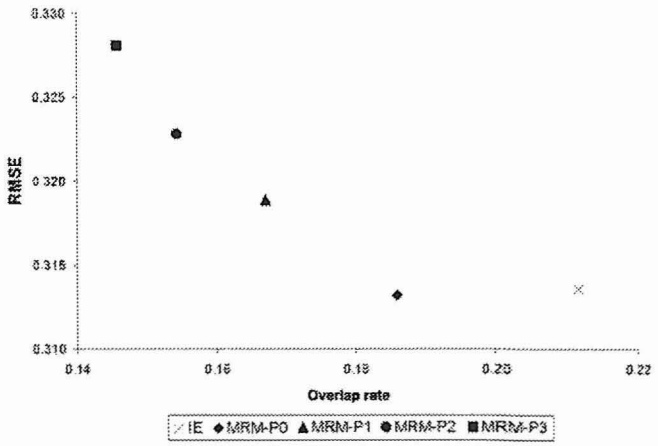
Simulation Study 2

Method

The method of this second study is similar to that of the first except in certain aspects. It used an item bank for assessing knowledge of English grammar, eCAT (Olea et al., 2004), used in human resources contexts for personnel selection and promotion. The bank has 197 items. This small size means that security issues are especially relevant for eCAT. The mean, standard deviation,

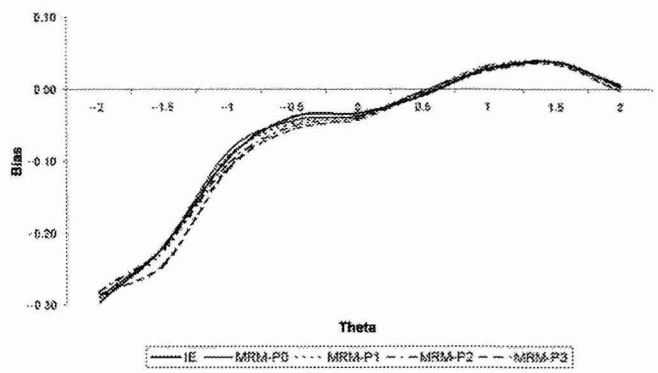
Volume 33 Number 1 January 2009

Figure 11
Overlap Rate and RMSE for the Operative Bank



Note: IE = item-eligibility method; RMSE = root mean square error.

Figure 12
Bias According to θ Values for the Operative Bank



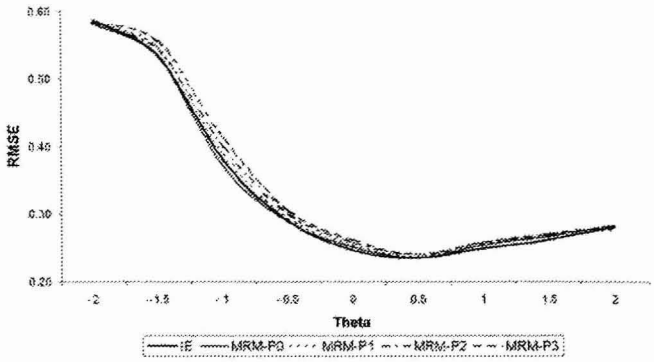
Note: IE = item-eligibility method.

maximum, and minimum for the *a*, *b*, and *c* parameters were (1.3, 0.32, 2.2, 0.43), (0.23, 1, 3.42, −2.71), and (0.21, 0.03, 0.29, 0.11), respectively. Test length was fixed at 20 items. Although in practice the test length depends on the needs of the companies that use it, this length is the one usually chosen. The maximum exposure rate was fixed at 0.25, as in the first study.

Results

Only the results for the exposure rates, overlap rates, RMSE, and bias at the end of the test are shown as these are the relevant data in a practical context.

Figure 13
RMSE According to θ Values for the Operative Bank



Note: IE = item-eligibility method; RMSE = root mean square error.

Figure 10 shows the exposure rate for the different methods. Basically, the results of Study 1 are replicated. Although with the item-eligibility method a considerable part of the item bank is never used, with the MRM method there is no item with an exposure rate equal to zero. Also, the MRM method reduces the proportion of items with an exposure rate close to the limit rate. These effects are more marked as the value of the acceleration parameter is increased.

Figure 11 shows the overlap rate and RMSE values. In accordance with that seen with the exposure rates, the item-eligibility method has the highest overlap rate. The greater the value of the acceleration parameter, the lower is the overlap rate. The differences in RMSE between methods are small, never greater than 0.015. The RMSE for the item-eligibility method and for the MRM method with an acceleration parameter equal to zero are the same. Accuracy correlates negatively with the acceleration parameter.

The differences in bias can be seen in Figure 12. As in Study 1, the differences between methods are mainly found in negative trait levels. The higher the acceleration parameter, the greater is the bias. The bias of the item-eligibility method and the bias of the MRM method with an acceleration parameter equal to 0 are indistinguishable. The same results are found for RMSE, which is shown in Figure 13.

The results found with eCAT (Olea et al., 2004) are mainly the same as those obtained with randomly generated item banks in Study 1. Compared with defining just a single maximum exposure rate, defining multiple exposure rates considerably improves the security of the item bank with either no or only minor decrements in measurement accuracy.

Conclusions

As noted above, if the examinees know some of the items before they take the test, the validity of the test is adversely affected. To reduce the occurrence of this problem, it is important to reduce the variance of item exposure rates and thus the overlap rate between examinees (Chen et al., 2003).

The approach most widely used in CAT has been to impose a maximum exposure rate that no item should exceed. To do so, control parameters are calculated that determine the probability of

an item's being administered once it has been selected, or the probability of its being eligible. Various methods have been proposed for calculating control parameters (Revuelta & Ponsoda, 1998; Sympson & Hetter, 1985; van der Linden & Veldkamp, 2004). However, several problems arise with this form of improved test security. First, although these methods are effective in eliminating overexposure, they have almost no impact on increased usage of underexposed items. Second, with these methods, when the maximum Fisher information selection rule is used, the quality of items selected decreases as the test progresses (Revuelta & Ponsoda, 1998).

This article presents an option for controlling exposure rates. Rather than defining only a single exposure rate as a limit, as many maximum exposure rates as items will be administered are marked. At the beginning of the test, the maximum exposure rate will be close to the minimum possible, increasing as the test progresses. The approach proposed involves only small modifications to the other method described, the item-eligibility method. A comparison of the performance of this new method with the item-eligibility method reveals that MRM clearly improves item bank security. Moreover, for the values of the acceleration parameter tested there is no relevant difference in accuracy between the MRM and the item-eligibility method. With randomly generated item banks, the overlap rate obtained with the item-eligibility method could be reduced by 40% while increasing RMSE by less than 0.01, when the acceleration parameter was set at 2. With eCAT, if an increment of 0.01 in RMSE is considered tolerable, the reduction in the overlap rate is 27%, with an acceleration parameter equal to 2.

Given the advantages of the MRM method, it can be considered the advisable option for controlling maximum exposure rates in CATs as it involves a more balanced usage of the item bank and delays possible distortion of trait estimation due to security problems with either no or only slight decrements in measurement accuracy.

References

Barrada, J. R., Olea, J., & Ponsoda, V. (2007). Methods for restricting maximum exposure rate in computerized adaptative testing. *Methodology, 3*, 14-23.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2006). Estrategias de selección de ítems en un test adaptativo informatizado para la evaluación del inglés escrito [Item selection rules in a computerized adaptive test for the assessment of written English]. *Psicothema, 18*, 828-834.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (in press). Incorporating randomness to the Fisher information for improving item exposure control in CATs. *British Journal of Mathematical and Statistical Psychology.*

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Chang, H. H. (2004). Understanding computerized adaptive testing—From Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 117-133). Thousand Oaks, CA: Sage.

Chang, H. H., & Ying, Z. (1999). α-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

Chang, H. H., & Ying, Z. (in press). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika.*

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*, 129-145.

Chen, S. Y., & Doong, S. H. (2008). Predicting item exposure parameters in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 61*, 75-91.

Chen, S. Y., Lei, P. W., & Liao, W. H. (in press). Controlling item exposure and test overlap on the fly in

computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*.

Hau, K. T., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38*, 249-266.

Li, Y. H., & Schafer, W. D. (2005). Increasing the homogeneity of CAT's item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. *Journal of Educational Measurement, 42*, 245-269.

Olea, J., Abad, F. J., Ponsoda, V., & Ximénez, M. C. (2004). Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: Diseño y comprobaciones psicométricas [A computerized adaptive test for the assessment of written English: Design and psychometric properties]. *Psicothema, 16*, 519-525.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Segall, D. O. (2004). Computerized adaptive testing. In K. Kempf-Lenard (Ed.), *The Encyclopedia of Social Measurement* (pp. 429-438). San Diego, CA: Academic Press.

Stocking, M. L., & Lewis, C. L. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57-75.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W. J. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 28*, 249-265.

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42*, 283-302.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273-291.

van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32*, 398-418.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17-27.

Acknowledgments

This research was partly supported by a grant from the Spanish Ministry of Education and Science (project SEJ2006-08313/PSIC).

Author's Address

Address correspondence to Juan Ramón Barrada, Facultad de Psicología, Universidad Autónoma de Barcelona, 08193 Bellaterra, Spain; e-mail: juanramon.barrada@uab.es.

Capítulo 4.
Incorporando aleatoriedad a la información de Fisher para la mejora del control de la exposición en TAIs

Artículo publicado. La referencia del mismo es:

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness to the Fisher information for improving item exposure control in CATs. *British Journal of Mathematical and Statistical Psychology, 61*, 493-513.

British Journal of Mathematical and Statistical Psychology (2008), 61, 493–513

The British Psychological Society

www.bpsjournals.co.uk

Incorporating randomness in the Fisher information for improving item-exposure control in CATs

Juan Ramón Barrada[1]*, Julio Olea[2], Vicente Ponsoda[2] and Francisco José Abad[2]

[1]Facultad de Psicología, Universidad Autónoma de Barcelona, Barcelona, Spain
[2]Facultad de Psicología, Universidad Autónoma de Madrid, Madrid, Spain

The most commonly employed item selection rule in a computerized adaptive test (CAT) is that of selecting the item with the maximum Fisher information for the estimated trait level. This means a highly unbalanced distribution of item-exposure rates, a high overlap rate among examinees and, for item bank management, strong pressure to replace items with a high discrimination parameter in the bank. An alternative for mitigating these problems involves, at the beginning of the test, basing item selection mainly on randomness. As the test progresses, the weight of information in the selection increases. In the present work we study, for two selection rules, the progressive methods (Revuelta & Ponsoda, 1998) and the proportional method (Segall, 2004a), different functions that define the weight of the random component according to the position in the test of the item to be administered. The functions were tested in simulated item banks and in an operative bank. We found that both the progressive and the proportional methods tolerate a high weight of the random component with minimal or zero loss of accuracy, while bank security and maintenance are improved.

1. Introduction

The basic objective of a computerized adaptive test (CAT) is to make accurate estimations of the trait levels of examinees through the presentation of a small number of items (Ponsoda & Olea, 2003; Segall, 2004b). Thus, with the different item selection rules proposed, the aim is to select the item that contributes most to reducing uncertainty about the true trait level of each examinee at each moment. In high-stakes tests, problems may emerge related to the content leakage of part of the item bank on which the test is based (Chang, 2004). Therefore, when item selection rules are

*Correspondence should be addressed to Juan Ramón Barrada, Facultad de Psicología, Universidad Autónoma de Barcelona, 08193 Bellaterra, Spain (e-mail: juanramon.barrada@uab.es).

DOI:10.1348/000711007X230937

proposed, both the measurement accuracy they provide and their capacity for maintaining bank security are usually examined. Another aspect to take into account, in order to make a CAT economically viable, is the ease of maintaining the bank.

In CATs, the most commonly employed item selection rule is the selection of the item that provides the most Fisher information for the estimated trait level ($\hat{\theta}$) at that point in the test (van der Linden & Pashley, 2000). The information for item i is given by

$$I_i(\theta) = \frac{[p_i'(\theta)]^2}{p_i(\theta)[1 - p_i(\theta)]} \tag{1}$$

where $p_i(\theta)$ is the probability of correct response to item i given trait level θ and $p_i'(\theta)$ is the first derivative of $p_i(\theta)$. In the three-parameter logistic model, $p_i(\theta)$ is calculated as

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}} \tag{2}$$

where a_i is the discrimination parameter, b_i is the localization parameter, and c_i is the pseudoguessing parameter, all of item i.

Defining event S_i as the selection of item i, this rule (point Fisher information, PFI), can be described as

$$\max_i I_i(\hat{\theta}) \rightarrow S_i. \tag{3}$$

Various authors have questioned the appropriateness of using this rule in a CAT. On the one hand, it does not seem to be the most efficient rule in terms of measurement accuracy (Chang & Ying, 1996; van der Linden, 1998; Veerkamp & Berger, 1997), especially when the number of items administered is small. As a CAT may normally contain over 10 items, the alternative rules to PFI that have been developed for improving measurement efficiency appear to offer fairly limited improvements (Chen & Ankenmann, 2004; Chen, Ankenmann, & Chang, 2000).

Furthermore, when PFI is used as a selection rule, the distribution of the exposure rate of the items is highly uneven, with some items administered to a high proportion of examinees and others never used. In general, the items with a high exposure rate will be those with a high value of the a parameter. By way of an example, in the work by Li and Schafer (2005), the correlation between the exposure rate and the value of the a parameter of the items was equal to .6.

This tendency signifies two problems for the functioning of a CAT: risks to item bank security and difficulty of maintenance. Chen, Ankenmann, and Spray (2003) have shown that the greater the variance in the exposure rates of the items, the greater the overlap rate between examinees. Overlap rate is the mean proportion of items that two examinees share (Way, 1998). The higher it is, the greater the risk to security, since an examinee receiving information from another of the items presented to him or her can lead to greater distortion in the estimation of the trait level.

In order to maintain the bank, it is necessary to withdraw certain items that may have become publicly known. The withdrawn items will be those with a high exposure rate, and therefore, in general, those with high discriminative capacity. If the new items to be incorporated in the bank follow the same distribution for parameter a as the items making up the bank, it is to be expected that the new items will have, as a mean, lower discriminative capacity than the items to be replaced. This would imply either a progressive deterioration in the quality of the bank or an increase in the costs of the CAT,

since it would be necessary to calibrate more than one item for each item withdrawn, until one with a similar a parameter is obtained. The greater the distance between the distribution of the a parameter of the withdrawn items and the distribution in the bank of that parameter, the greater the problem will be (Hau & Chang, 2001).

Various methods of exposure control have been proposed. In the present work, we focus on two methods, the progressive (Revuelta & Ponsoda, 1998) and the proportional (Segall, 2004a), introducing modifications to them that can improve security and facilitate the maintenance of the bank, with little or no loss of measurement accuracy. With this purpose in mind, we first examine those which up to now have been the most commonly used methods of exposure control in CATs, before moving on to the methods that we wish to evaluate.

2. Methods for exposure control

Various alternatives have been proposed for improving item-exposure control. The approach that has attracted the most attention has been that of fixing a maximum exposure rate ($r^{\max}$) which no item should exceed (Revuelta & Ponsoda, 1998; Sympson & Hetter, 1985; van der Linden & Veldkamp, 2004). Chen and Lei (2005) have proposed a method that simultaneously controls maximum exposure rate and overlap rate.

Despite its limitations (Barrada, Olea, & Ponsoda, 2007; van der Linden, 2003), the most commonly employed method for restricting maximum exposure rate is the Sympson–Hetter method. When this method is used, two different events are defined: item i is selected by the item selection rule (S_i); and item i is administered (A_i). With the Sympson–Hetter method, selection of item i does not imply its administration. The probability of administration of item i conditional on selection of that item, $[P(A_i|S_i)]$, is the parameter that, appropriately defined, would allow fulfilment of the criterion that no item-exposure rate is above $r^{\max}$. When an item is selected, a random number belonging to U(0, 1) is generated, and only if this number is lower than $P(A_i|S_i)$ is that item administered. In the opposite case, the item is not administered and is marked as non-selectable for that examinee.

As an item cannot be administered if it has not been selected, it follows that

$$P(A_i) = P(A_i|S_i)P(S_i). \tag{4}$$

The values of $P(A_i|S_i)$ are obtained through a series of simulation cycles. The $P(A_i|S_i)$ parameters for cycle $z + 1$ derive from making $P(A_i)$ equal to $r^{\max}$ in equation (4) and setting the limitation that the maximum permitted value of the $P(A_i|S_i)$ parameter is 1.

$$P(A_i|S_i)^{(z+1)} = \begin{cases} 1, & \text{if } P^{(z)}(S_i) \leq r^{\max}. \\ r^{\max}/P^{(z)}(S_i), & \text{if } P^{(z)}(S_i) > r^{\max}. \end{cases} \tag{5}$$

When the value of the maximum exposure rate is stabilized or a fixed number of cycles is reached, the simulation process for establishing the $P(A_i|S_i)$ parameters is stopped.

The methods for limiting the maximum exposure rate involve two main disadvantages: (a) exposure control is achieved at the cost of a loss of measurement accuracy, to the point that some authors actually assume this balance between the two variables to be necessary (Stocking & Lewis, 2000); (b) in general, these methods increase the presentation of the items with an a parameter slightly below the a parameter of the

overexposed items, so that they barely increase the exposure rate of the underexposed items (i.e. items with null or very small exposure rates) or bring the distribution of items to be withdrawn close to the original distribution of the bank. The lower the value of r^{max}, the more marked will be the problem described in (a), and the less marked that in (b).

Another approach to controlling the exposure rate is that of stratified methods (Chang & Ying, 1999). In these, both the item selection rule and the definition of the available bank in each phase of the test are changed. Items are no longer chosen according to the Fisher information function, but rather according to the difference, in absolute value, between the *b* parameter of the item and the estimated trait level. The sub-bank from which items can be selected, at the beginning of the test, will be made up of the items with a low value of the *a* parameter. As the test progresses, the discriminative capacity of the available items increases. This strategy, or one of its variants (Barrada, Mazuela, & Olea, 2006), highly effective in the control of exposure, brings the *a* parameter distribution of the items used very close to the distribution of this parameter in the complete bank (Hau & Chang, 2001). The drawback is that, in achieving this, there is a loss of measurement accuracy.

3. Incorporating a random element in the Fisher information

We now describe two methods that, based on the Fisher information function, incorporate a random element that permits an increase in bank security with a minimal loss of accuracy: the progressive method (Revuelta & Ponsoda, 1998) and the proportional method (Segall, 2004a). The weight of the random component will be greater, the closer we are to the start of the test, when estimation of trait level is more unstable and measurement error is greater. The speed with which the impact of the random element decreases in these methods will mark the capacity for exposure control.

3.1. Progressive method

In the progressive method (PG), proposed by Revuelta and Ponsoda (1998), the item selected is that with maximum value for the sum of a random component, of major importance at the start of the test, and the Fisher information, of major importance towards the end of the test. For selection of the *k*th item, *k* being the position of the item in the test, PG determine

$$\max_i [(1 - W_k)R_i + W_k I_i(\hat{\theta})] \rightarrow S_i \tag{6}$$

where W_k is the proportion of the selection criterion due to the information function and R_i is a random number in the interval $[0, \max_i I_i(\hat{\theta})]$.

In the original proposal, W_k is defined in a linear fashion where the minimum value is equal to 0 and the maximum is $(K - 1)/K$, K being the number of items to be administered:

$$W_k = \frac{K - 1}{K}. \tag{7}$$

Revuelta and Ponsoda (1998) showed that this method leads to an improvement in exposure control with slight loss of measurement accuracy. The improvement in security with low impact on the quality of the estimation is due to the fact that, in comparison with PFI, although at the beginning of the test the items offer much less information, at the end of the test PG offers more informative items than PFI, since for

PFI the highly discriminative items have run out. Thus, the information accumulated by the two methods is not actually very different.

Eggen (2001) introduced two modifications to the original PG method. First, he proposed the simultaneous application of the PG and Sympson-Hetter methods as a partial solution for overexposure and underexposure control problems. Second, he suggested applying the PG method and reducing the random component contribution to specific parts of the test. For example, the PG method might be applied to just the first half of the test whereas the items on the second half would be selected by PFI.

3.2. Proportional method

When PFI is applied, the items that are not, for any possible trait level, among the K items with highest value of the information function will never be selected, regardless of the information they might provide. In order to address this disadvantage and improve exposure control, Segall (2004a) proposed using a different item selection method from PFI, which we shall call the proportional method (PP). In Segall's proposal, the information provided by each item in a bank of size N serves to determine the probability of its selection:

$$p(S_i) = \frac{V_i}{\sum_{i=1}^{N} V_i}. \tag{8}$$

By doing this, it is guaranteed that: (a) the sum of the probabilities of selection for the N items is equal to 1; (b) no single item has a probability of selection equal to zero.

Even though in Segall's paper V_i is the expected variance (van der Linden, 1998), the application with the Fisher information is direct in making

$$V_i = I_i(\hat{\theta}). \tag{9}$$

Once the probabilities for all the items have been determined, a cumulative distribution is constructed with the $p(S_i)$ values. A random number R, in the interval $[0, 1]$, is generated. By means of this number the item to be selected is determined:

$$\sum_{j=1}^{i-1} p(S_j) < R \leq \sum_{j=1}^{i} p(S_j) \rightarrow S_i. \tag{10}$$

Thus, for example, with three items with a, b and c parameters equal to $(2, \hat{\theta}, 0)$, $(1.9, \hat{\theta}, 0)$ and $(1, \hat{\theta}, 0)$, the probability of selecting each of them using equations (8)-(10) is, respectively, .46, .42 and .12. The first item will be selected if R is less than or equal to .46; the second item will be selected if R is less than or equal to .88 (.46 + .42) and greater than .46; and the third item will be selected if R is greater than .88.

The PP method considerably reduces the variance of the exposure rates but basically provides the same item selection probabilities throughout the test. While it appears advantageous to favour the selection of items with low a parameters at the beginning of the test, these should have low probability of selection in the final stages of the test. To achieve this, we propose using the information raised to a power (P_k) that increases as the test progresses.

$$V_i = I_i(\hat{\theta})^{P_k}. \tag{11}$$

If we make P_k equal to 0, all the V_i values will be equal to 1. Thus, the selection will be random. With P_k equal to 1, the method would be applied as proposed by Segall. For the same three fictitious items as above, with P_k equal to 5, the probability of selection of

each item (ordered by its a parameter) is .63, .37, and .00. With P_k equal to 20, the probabilities would be .89, .11, and .00. The higher the value of P_k, the greater will be the similarity between item selection with PP and selection by means of PFI.

3.3. Functions for W_k and P_k

The PG and PP methods use the W_k and P_k functions, which enable them to define the weight of chance in the selection of items for each test item position. At the beginning of the test, as indicated by Revuelta and Ponsoda (1998) and Li and Schafer (2005), there appears to be no problem in selecting items at random; at the end of the test, our aim is for these two rules to work in a similar way to PFI.

There are several possible functions that would permit us to relate the position of the item in the test (k) to the value of W_k and P_k. We have opted for the functions

$$W_k = \begin{cases} 0, & \text{if } k = 1, \\ \dfrac{\sum_{b=2}^{k} (b-1)^s}{\sum_{b=2}^{k} (b-1)^s}, & \text{if } k \neq 1, \end{cases} \tag{12}$$

$$P_k = \begin{cases} 0, & \text{if } k = 1, \\ \dfrac{K\sum_{b=2}^{k} (b-1)^s}{\sum_{b=2}^{k} (b-1)^s}, & \text{if } k \neq 1, \end{cases} \tag{13}$$

where b is a bound variable, used only for calculations, and t and s are parameters that permit control of the speed with which W_k and P_k move away from 0 to reach the maximum. We shall call them acceleration parameters.

These two increasing functions imply random selection at the beginning of the test. For the selection of the last item, W_k is equal to 1 and P_k is equal to K. When W_k equals 1, PG and PFI are equivalent rules. With P_k equal to K, for the sizes of K normally employed in a CAT, the probability of selecting an item that offers low information will be negligible.

For negative values of the acceleration parameters, there will be a rapid switch from random selection to selection based on the information. For a value of the acceleration parameters equal to 0, this transition is linear. For positive values of these parameters, the importance of the random component will be maintained for longer the higher the value of t or of s. Negative acceleration parameter values in the PG method would lead to a reduction of the importance of the random component, when compared with Revuelta and Ponsoda's (1998) original proposal, as in the study conducted by Eggen (2001). The effect of changes in the acceleration parameters can be seen in Figure 1.

In order to compare the functioning of PFI, PG, and PP we carried out two simulation studies. In the first of these, we used banks with theoretical distributions of parameters. In the second, we used the estimated parameters of a currently operative bank (Olea, Abad, Ponsoda, & Ximénez, 2004).

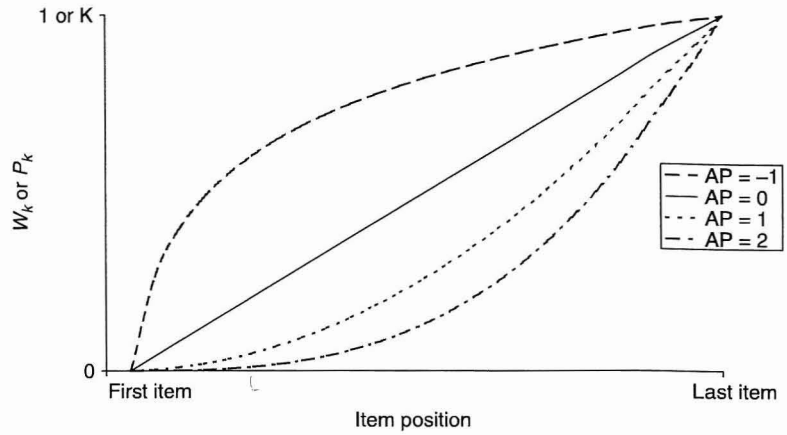


Figure 1. Item position in the test and W_k (proportion of the selection criterion due to the information function in the PG method) or P_k (power to which the information is raised in the PP method) for four different acceleration parameters (AP).

4. STUDY 1

4.1. Method

4.1.1. Item banks

Ten item banks of 500 items were generated. The distributions for the parameters were: $a \sim N(1.2, 0.25)$; $b \sim N(0, 1)$; $c \sim N(0.25, 0.02)$.

4.1.2. Trait level of the simulees, starting rule, and test length

The trait level of the simulees was randomly generated from a population $N(0, 1)$. For each item bank, 5,000 simulees were sampled. The starting $\hat{\theta}$ was chosen at random from the interval $(-0.5, 0.5)$. There were two different test lengths: 25 and 40 items.

4.1.3. Estimation/assignment of trait level

Maximum-likelihood estimation cannot be carried out with real numbers when there is a constant response pattern, all correct or all incorrect responses. In order to avoid this, until there was at least one correct and one incorrect response, $\hat{\theta}$ was assigned using the method proposed by Dodd (1990). When all the responses were correct, $\hat{\theta}$ was increased by $(b_{\max} - \hat{\theta})/2$. If all the responses were incorrect, $\hat{\theta}$ was reduced by $(\hat{\theta} - b_{\min})/2$. In these formulae $b_{\max}$ and $b_{\min}$ refer to the highest and lowest b parameters, respectively, of the entire item bank. When the constant pattern was broken or when the test was finished, maximum-likelihood estimation was applied, with the restriction that $\hat{\theta}$ had to be in the interval $[-4, 4]$.

4.1.4. Restrictions on $r^{\max}$

We simulated two different conditions. In the first, no restriction on $r^{\max}$ was applied. In the second, $r^{\max}$ was set equal to .25. To do so, the Sympson-Hetter method was

employed. The $P(A_i|S_i)$ parameters used were those generated after the 20th cycle, calculated with equation (5).

4.1.5. Content balancing

Different examinees should face similar content during the test (van der Linden, 2005). Two testing conditions were simulated. In the first, no content constraints were imposed. In the second, the items of the bank were randomly distributed in five different content categories, each with the same number of items. We imposed the restriction that one fifth of the items in the test should be selected from each content area. To do so, a modification of the method proposed by Kingsbury and Zara (1989) was applied. The number of items not yet selected from each category divided by the number of items not yet presented is computed at each item position. A cumulative distribution is constructed with these values, following a multinomial distribution. A random number from U(0, 1) determines the content area from which the next item will be selected. A similar approach was used by Leung, Chang, and Hau (2003). In the content balancing conditions, $I(\hat{\theta})$ was computed only for the items belonging to the content area chosen for item selection.

4.1.6. W_k and P_k values

These values were computed for each item position in the test as shown in equations (12) and (13). We employed four different values for the acceleration parameters: −1, 0, 1 and 2. We also evaluated Segall's original proposal (2004a), where P_k is held constant and equal to 1. We have called this proposal S04. Because of the minimal differences in the definition of W_k between equations (7) and (12), with the acceleration parameter equal to zero, we did not specifically test Revuelta and Ponsoda's original proposal (1998). For clarity of exposition, when we speak hereafter about the PP method, we are referring only to the conditions with the P_k values defined according to equation (13), and not to the S04 method, even though it is a proportional method.

4.1.7. Performance measures

Four dependent variables were used for the comparison between methods: (a) exposure rates of the items; (b) root mean square error (RMSE) for the accuracy, as shown in equation (14); (c) overlap rate to measure the security risk, as defined in equation (15) (Chen *et al.*, 2003); and (d) the mean *a* parameter value of the items administered, as information about the item bank maintenance problems.

The RMSE was

$$\text{RMSE} = \left(\sum_{g=1}^{r} (\hat{\theta}_g - \theta_g)^2 / r \right)^{1/2} \tag{14}$$

where r is the number of simulees.

The overlap rate was

$$\hat{T} = \frac{N}{K} S^2_{P(A)} + \frac{K}{N} \tag{15}$$

where $\hat{T}$ is the large-sample approximation of the overlap rate (Chen *et al.*, 2003), N is the bank size, and $S^2_{P(A)}$ is the variance of the item-exposure rates.

Summarizing the design of Study 1, we manipulated the following variables: test length, restrictions on r^{max}, content balancing, and item selection rules.

4.2. Results

Tables 1 and 2 show the exposure rates of the items for the different selection rules according to the different simulation conditions. First, we will describe the overall results according to the selection rules. Then, we will indicate the effect of the variables manipulated.

As expected, with PFI a large part of the bank is not presented to any examinee, ranging from 35% to 62%. If we set a rate over .25 as the overexposure criterion, with the PFI rule, from 4% to 12% of the items could be considered overexposed. The PG and PP rules succeed in reducing the variance of the exposure rates. With these, no item has a rate equal to 0 and the proportion of overexposed items is reduced. These effects increase, the higher the value of the acceleration parameter. The lower the acceleration parameter, the more similar are the exposure rate distributions to those obtained with PFI. Similar results were obtained by Eggen (2001) when the PG method was applied to just a part of the test. In the case of S04, exposure distribution is fairly similar to that which could be expected with random selection.

Logically, the Sympson–Hetter method succeeds in reducing overexposure, but the method was unable to fix all of the exposure rates below r^{max} (van der Linden, 2003). As noted before, restrictions on r^{max} slightly reduce the proportion of underexposed items. Meeting some content balance requirements, probably because the method used for this also included a random component, reduced the number of underexposed and overexposed items. This effect was more marked for PG than for PP. Increasing the test length implies increasing the mean exposure rate of the items. When passing from 25 to 40 items to be administered, the number of underexposed items decreased and the number of overexposed items increased.

Figures 2 and 3 show the result obtained with the different rules, for both RMSE and the overlap rate. The pattern of results is basically the same for the two different test lengths. Following what was presented in Tables 1 and 2, both with PG and with PP, in comparison with PFI, the overlap rate decreases. This reduction is more marked the higher the value of the acceleration parameter. With S04, the overlap rate is near the minimum possible, given the test length and item bank size. Incorporating content balancing leads to a small reduction in the overlap rate both for PFI and PP. This effect of content balancing is much greater with PG. Imposing r^{max} equal to .25 also reduces the overlap rate. This impact will be greater the higher the proportion of items with an exposure rate over .25 when no restriction is applied.

In terms of accuracy, S04 is the method with the higher RMSE. In general, both for PG and PP, as the value of the acceleration parameter increases, the RMSE increases. The differences in RMSE are greater when we move from s or t values of 1 to 2 than when we move from values of -1 to 0.

PFI is not always the selection rule that leads to the higher accuracy. Furthermore, when PG or PP improves accuracy, in comparison with PFI, the difference in RMSE is very small. Controlling for content balance, for PFI, PP and S04, has a very small impact on accuracy. For PG, the differences between the condition without content balance

Table 1. Proportion of items with different item-exposure rates in Study 1 with a test length of 25 items for the different item selection rules and the acceleration parameters. In this and the following tables and figures, the number after the ^ symbol indicates the value of the acceleration parameter. NCB refers to the no content balancing condition and CB refers to the condition with content balancing

	Exposure rate	PFI	PG^(−1)	PG^0	PG^1	PG^2	PP^(−1)	PP^0	PP^1	PP^2	S04
$r^{max} = 1$ NCB	0	.62	0	0	0	0	0	0	0	0	0
	(0, .02]	.08	.70	.66	.53	.35	.68	.64	.47	.13	.01
	(.02, .05]	.05	.05	.09	.22	.41	.07	.10	.28	.64	.50
	(.05, .1]	.07	.06	.07	.10	.12	.07	.08	.10	.12	.49
	(.1, .15]	.05	.05	.05	.05	.05	.05	.06	.05	.04	0
	(.15, .2]	.03	.04	.04	.04	.03	.04	.04	.04	.03	0
	(.2, .25]	.03	.03	.03	.03	.02	.03	.03	.02	.02	0
	(.25, .3]	.02	.02	.02	.02	.01	.02	.02	.02	.01	0
	(.3, .4]	.02	.03	.02	.02	.01	.02	.02	.01	.01	0
	(.4, .5]	.01	.01	.01	0	0	.01	0	0	0	0
	(.5, 1]	.01	0	0	0	0	0	0	0	0	0
$r^{max} = 1$ CB	0	.57	0	0	0	0	0	0	0	0	0
	(0, .02]	.11	.65	.51	.27	.04	.66	.60	.44	.12	.01
	(.02, .05]	.06	.08	.20	.44	.67	.08	.12	.28	.61	.51
	(.05, .1]	.08	.09	.12	.17	.19	.09	.10	.13	.14	.48
	(.1, .15]	.05	.06	.06	.07	.06	.06	.07	.06	.06	0
	(.15, .2]	.03	.04	.04	.03	.03	.04	.04	.04	.03	0
	(.2, .25]	.03	.03	.02	.02	.01	.03	.03	.02	.02	0
	(.25, .3]	.02	.02	.02	.01	0	.02	.02	.01	.01	0
	(.3, .4]	.03	.03	.01	0	0	.03	.02	.01	0	0
	(.4, .5]	.01	.01	0	0	0	.01	0	0	0	0
	(.5, 1]	0	0	0	0	0	0	0	0	0	0

Table 1. (*Continued*)

	Exposure rate	PFI	PG^(−1)	PG^0	PG^1	PG^2	PP^(−1)	PP^0	PP^1	PP^2	S04
$r^{max} = .25$ NCB	0	.57	0	0	0	0	0	0	0	0	0
	(0, .02]	.09	.68	.64	.52	.34	.65	.62	.46	.13	.01
	(.02, .05]	.05	.06	.09	.22	.4	.07	.11	.29	.63	.50
	(.05, .1]	.07	.06	.07	.10	.12	.07	.08	.10	.12	.49
	(.1, .15]	.05	.05	.06	.05	.05	.05	.06	.05	.04	0
	(.15, .2]	.04	.04	.04	.04	.03	.04	.04	.04	.03	0
	(.2, .25]	.08	.08	.07	.05	.04	.07	.06	.04	.03	0
	(.25, .3]	.05	.04	.03	.02	.02	.04	.03	.02	.01	0
	(.3, 1]	0	0	0	0	0	0	0	0	0	0
$r^{max} = .25$ CB	0	.51	0	0	0	0	0	0	0	0	0
	(0, .02]	.14	.63	.51	.26	.04	.63	.59	.43	.12	.01
	(.02, .05]	.06	.08	.20	.45	.66	.09	.12	.28	.61	.50
	(.05, .1]	.08	.08	.13	.17	.19	.08	.10	.13	.15	.49
	(.1, .15]	.05	.06	.06	.06	.06	.06	.07	.07	.06	0
	(.15, .2]	.04	.04	.05	.03	.03	.04	.05	.04	.03	0
	(.2, .25]	.08	.06	.04	.02	.01	.07	.05	.04	.03	0
	(.25, .3]	.04	.03	.02	.01	0	.03	.02	.01	.01	0
	(.3, 1]	0	0	0	0	0	0	0	0	0	0

Table 2. Proportion of items with different item-exposure rates in Study 1 with a test length of 40 items for the different item selection rules and acceleration parameters

	Exposure rate	PFI	PG^(−1)	PG^0	PG^1	PG^2	PP^(−1)	PP^0	PP^1	PP^2	S04
$r^{max} = 1$ NCB	0	.46	0	0	0	0	0	0	0	0	0
	(0, .02]	.09	.56	.54	.30	.02	.55	.53	.26	0	0
	(.02, .05]	.06	.06	.09	.32	.56	.07	.08	.37	.61	.10
	(.05, .1]	.09	.08	.08	.12	.19	.08	.09	.11	.15	.71
	(.1, .15]	.07	.07	.07	.07	.07	.07	.08	.07	.07	.19
	(.15, .2]	.06	.06	.06	.06	.05	.06	.06	.06	.05	0
	(.2, .25]	.05	.05	.05	.05	.05	.05	.05	.05	.04	0
	(.25, .3]	.04	.04	.05	.04	.03	.04	.05	.04	.03	0
	(.3, .4]	.05	.05	.05	.04	.03	.04	.05	.04	.03	0
	(.4, .5]	.01	.02	.01	.01	0	.02	.02	.01	0	0
	(.5, 1]	.01	0	0	0	0	.01	0	0	0	0
$r^{max} = 1$ CB	0	.43	0	0	0	0	0	0	0	0	0
	(0, .02]	.11	.53	.34	.01	0	.53	.5	.24	0	0
	(.02, 05]	.07	.08	.24	.5	.41	.08	.10	.36	.57	.10
	(.05, .1]	.09	.09	.14	.24	.36	.10	.10	.14	.18	.71
	(.1, .15]	.07	.08	.09	.10	.11	.07	.08	.08	.09	.19
	(.15, .2]	.06	.06	.06	.06	.06	.06	.07	.06	.06	0
	(.2, .25]	.05	.05	.05	.04	.03	.05	.05	.05	.04	0
	(.25, .3]	.04	.04	.04	.02	.02	.04	.04	.04	.03	0
	(.3, .4]	.05	.05	.03	.02	.01	.05	.05	.03	.02	0
	(.4, .5]	.02	.01	0	0	0	.02	.01	0	0	0
	(.5, 1]	.01	0	0	0	0	0	0	0	0	0

Table 2. (*Continued*)

	Exposure rate	PFI	PG^(−1)	PG^0	PG^1	PG^2	PP^(−1)	PP^0	PP^1	PP^2	S04
$r^{max} = .25$ NCB	0	.41	0	0	0	0	0	0	0	0	0
	(0, .02]	.10	.52	.50	.27	.01	.50	.50	.22	0	0
	(.02, .05]	.06	.06	.09	.32	.54	.07	.08	.38	.59	.09
	(.05, .1]	.08	.08	.08	.13	.19	.09	.09	.11	.16	.72
	(.1, .15]	.07	.07	.07	.07	.07	.07	.08	.07	.07	.19
	(.15, .2]	.06	.06	.06	.06	.05	.06	.06	.06	.05	0
	(.2, .25]	.13	.12	.13	.10	.08	.13	.12	.10	.09	0
	(.25, .3]	.08	.09	.07	.05	.04	.07	.07	.06	.04	0
	(.3, 1]	0	0	0	0	0	0	0	0	0	0
$r^{max} = .25$ CB	0	.35	0	0	0	0	0	0	0	0	0
	(0, .02]	.14	.49	.32	.01	0	.49	.47	.21	0	0
	(.02, .05]	.07	.08	.23	.49	.41	.08	.10	.36	.56	.09
	(.05, .1]	.09	.09	.15	.23	.35	.09	.10	.14	.18	.71
	(.1, .15]	.08	.08	.09	.11	.11	.08	.08	.09	.09	.19
	(.15, .2]	.06	.06	.07	.07	.06	.06	.07	.06	.06	0
	(.2, .25]	.13	.13	.10	.07	.04	.13	.12	.09	.07	0
	(.25, .3]	.08	.07	.05	.02	.02	.07	.06	.05	.04	0
	(.3, 1]	0	0	0	0	0	0	0	0	0	0

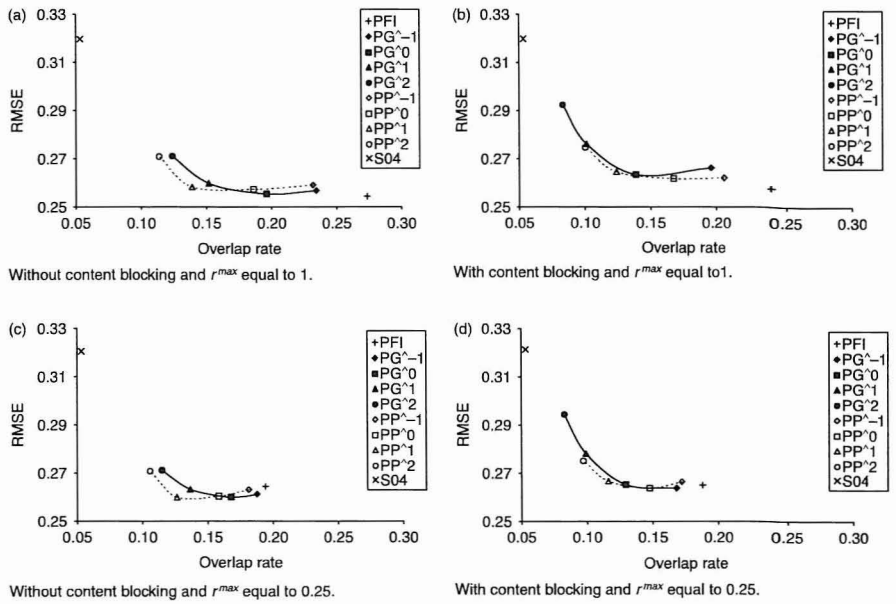


Figure 2. RMSE and test overlap for the theoretical item banks with a test length of 25 items.

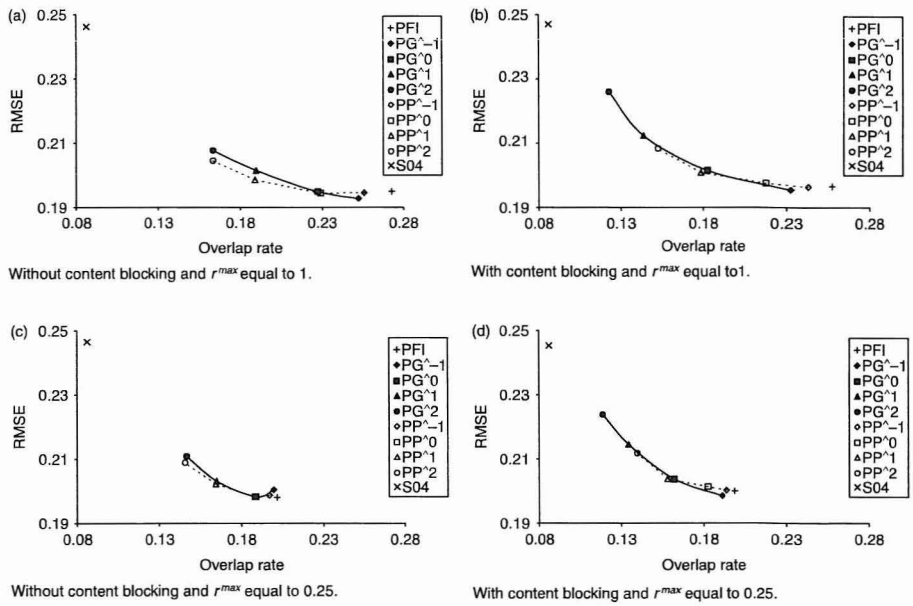


Figure 3. RMSE and test overlap for the theoretical item banks with a test length of 40 items.

and with content balance are higher. For all the methods, the impact on RMSE of restricting the maximum exposure rate is small.

Figures 2 and 3 allow the simultaneous visualization of both RMSE and overlap. In our opinion, this kind of figure is one of the most useful when comparing different item

selection rules. As expected, S04 is at the extreme of high security with low accuracy, and PFI at the other extreme: lower security with an accuracy that is maximum or near maximum. It can be seen how, with PG and PP, it is possible to improve the item bank security with little or no loss of measurement accuracy. For all the conditions simulated, it is possible to find one alternative method to PFI with an equivalent RMSE, but with a lower overlap rate. It can be seen how the manipulation of the acceleration parameters makes it possible to control the degree of security we obtain. In six out of eight conditions, the lines defined by the PG and PP methods show the same pattern: for high overlap rate values, the method with a lower RMSE is PG; for lower overlap rate values, the most accurate rule is PP.

Tables 3 and 4 contain the mean of the *a* parameter of the items administered. The mean of this parameter in the bank was equal to 1.2. The highest value is obtained with PFI; the lowest with S04. As expected, the pattern of the items selected with all of the methods is greatly different from the distribution of items in the bank. The only exception is S04. With the PG and PP methods, increasing the acceleration parameter reduces the mean value of the *a* parameter. Increasing the test length, incorporating content blocking, and restricting r^{max} reduce the mean discrimination of the items presented. As explained above, the effect of applying the Sympson–Hetter method is fairly limited in making the distribution of the items administered more similar to the items in the bank.

4.3. Discussion

As expected, the methods described are effective in the improvement of exposure control, when compared with PFI. The method with the best exposure control is S04. In the PG and PP methods, the higher the acceleration parameter value, that is, the longer the time during which the random component plays an important role, the fewer the risks to bank security. We also succeeded in bringing the *a* parameter distribution of the items used closer to the distribution of the items in the bank, thus facilitating bank maintenance. The PP and PG methods also had a small impact on reducing overexposure, which was greater the higher the acceleration parameter. All of this can be achieved at the cost of a very small or no loss of

Table 3. Mean *a* parameter of the presented items in Study 1 with a test length of 25 items for the different item selection rules and the different acceleration parameters (the mean *a* value in the bank was 1.2)

	r^{max} = 1 NCB	r^{max} = 1 CB	r^{max} = .25 NCB	r^{max} = .25 CB
PFI	1.56	1.53	1.52	1.50
PG^(−1)	1.54	1.50	1.50	1.48
PG^0	1.51	1.45	1.49	1.44
PG^1	1.46	1.39	1.45	1.39
PG^2	1.42	1.35	1.41	1.35
PP^(−1)	1.54	1.51	1.50	1.49
PP^0	1.50	1.48	1.48	1.46
PP^1	1.44	1.42	1.43	1.41
PP^2	1.40	1.38	1.39	1.37
S04	1.24	1.24	1.24	1.24

Table 4. Mean *a* parameter of the presented items with a test length of 40 items for different item selection rules and different acceleration parameters in Study 1 (the mean *a* value in the bank was 1.2)

	$r^{max} = 1$ NCB	$r^{max} = 1$ CB	$r^{max} = .25$ NCB	$r^{max} = .25$ CB
PFI	1.50	1.48	1.45	1.44
PG^(−1)	1.48	1.46	1.44	1.43
PG^0	1.46	1.42	1.43	1.40
PG^1	1.43	1.37	1.40	1.36
PG^2	1.40	1.34	1.38	1.34
PP^(−1)	1.49	1.47	1.44	1.43
PP^0	1.47	1.45	1.43	1.42
PP^1	1.42	1.41	1.40	1.39
PP^2	1.39	1.38	1.37	1.36
S04	1.24	1.24	1.24	1.24

accuracy. In our simulation study, these results hold for different test lengths, with and without content blocking and with and without restrictions on r^{max}.

All these conclusions are limited to the representativeness of our randomly generated item banks. In Study 2, we wanted to evaluate whether these results hold in an operative item bank, with estimated item parameters and real content areas.

5. STUDY 2

5.1. Method

The method for this second study was similar to that of the first study, except in relation to some points to which we shall now refer.

For this study, we used an item bank for assessing knowledge of English grammar, the eCAT (Olea *et al.*, 2004). The bank is made up of 197 multiple-choice items, with four response categories, of which just one is correct. Four different content categories were defined in the construction of the item bank. The values of the mean and the standard deviation for the estimated *a*, *b*, and *c* parameters of the item bank and the categories can been seen in Table 5.

The length of the test was set at 20 items. Even though in practice the test length depends on the needs of the companies that use it, this length appears sufficient for ordinary purposes (Barrada, Olea, Ponsoda, & Abad, 2006; Olea *et al.*, 2004). The test is usually administered without content blocking requirements. We included a condition

Table 5. Parameter distribution across content areas in Study 2

		a		*b*		*c*	
	N	Mean	*SD*	Mean	*SD*	Mean	*SD*
Content 1	92	1.30	0.29	−0.05	0.88	0.21	0.03
Content 2	61	1.32	0.32	0.55	1.14	0.21	0.02
Content 3	28	1.36	0.36	0.41	0.51	0.21	0.02
Content 4	16	1.18	0.35	0.30	1.19	0.21	0.02
Total	197	1.30	0.32	0.23	1	0.21	0.03

with content balancing so as to make a more complete comparison with Study 1. The number of items to be selected from each content area was 9, 6, 3, and 2 for contents areas 1-4. We maintained the restriction on the maximum rate with which eCAT currently functions. We set r^{max} equal to .25 and, as in Study 1, we applied the Sympson-Hetter method, using the $P(A_i|S_i)$ parameters generated after the 20th cycle.

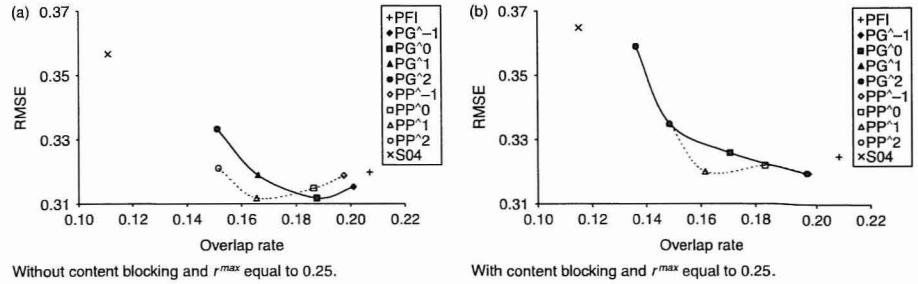
5.2. Results

The exposure rates of the items, as Table 6 shows, repeat the patterns obtained with the theoretical banks when the Sympson-Hetter method was applied. When PFI is used, depending on the absence (presence) content blocking, 24% (31%) of the bank is never used and more than 10% of the items have an exposure rate of over r^{max}. With the PG and PP rules, no item has an exposure rate equal to 0 and a smaller number of items have an exposure rate equal to or higher than r^{max}. The higher the value of the acceleration parameter, the greater the gain in exposure control. With S04, the distribution of the exposure rates is fairly similar to that which could be obtained through random selection.

Figure 4 shows the results for RMSE and overlap rate. As in Study 1, both PG and PP reduce the overlap rate, and this improvement is greater the higher the value of the acceleration parameter. In this study, as we saw in some conditions in Study 1, it is possible to obtain lower levels of RMSE with PG and PP for some values of the acceleration parameter. As mentioned earlier, this result is surprising, since we find that the alternative rules to PFI allow us to simultaneously improve measurement accuracy and bank security. We consider that this may be due to the fact that, in comparison to PFI, both PG and PP present fewer items with problems of overexposure, as we saw in Study 1. Thus, when we limit the maximum exposure rate, for PFI there will be a larger

Table 6. Proportion of items with different item-exposure rates in Study 2 for the different item selection rules and the different acceleration parameters

	Exposure rate	PFI	PG^(−1)	PG^0	PG^1	PG^2	PP^(−1)	PP^0	PP^1	PP^2	S04
NCB	0	.31	0	0	0	0	0	0	0	0	0
	(0,.02]	.08	.40	.32	0	0	.38	.34	0	0	0
	(.02,.05]	.08	.09	.15	.41	.28	.10	.12	.42	.27	.03
	(.05,.1]	.10	.10	.13	.23	.39	.10	.14	.21	.41	.45
	(.1,.15]	.08	.08	.08	.09	.11	.09	.08	.10	.11	.44
	(.15,.2]	.08	.06	.06	.07	.05	.07	.08	.06	.05	.08
	(.2,.25]	.15	.17	.17	.11	.12	.16	.16	.13	.11	0
	(.25,.3]	.13	.10	.09	.09	.05	.11	.09	.07	.05	0
	(.3,1]	0	0	0	0	0	0	0	0	0	0
CB	0	.24	0	0	0	0	0	0	0	0	0
	(0,.02]	.17	.37	.11	0	0	.38	.27	0	0	.01
	(.02,.05]	.06	.11	.33	.3	.12	.09	.19	.42	.28	.05
	(.05,.1]	.10	.11	.18	.33	.53	.11	.14	.20	.36	.51
	(.1,.15]	.08	.08	.09	.14	.17	.08	.08	.11	.14	.34
	(.15,.2]	.05	.06	.10	.10	.07	.06	.07	.09	.09	.09
	(.2,.25]	.19	.20	.12	.08	.08	.18	.17	.13	.09	.01
	(.25,.3]	.11	.08	.07	.05	.03	.10	.07	.05	.05	0
	(.3,1]	0	0	0	0	0	0	0	0	0	0



Without content blocking and r^{max} equal to 0.25. With content blocking and r^{max} equal to 0.25.

Figure 4. RMSE and test overlap for the operative item bank.

proportion of items that cannot be administered than for the other two methods. With the S04 method a security control near the maximum possible is reached, while the accuracy is made worse.

When comparing the condition of content blocking with the condition of no content blocking, the results show some differences. First, for all the conditions the RMSE is greater when imposing content balancing. And second, while in the no content blocking condition the PG method was better for some overlap levels, with content blocking the PG is never a better alternative than PP (at least, for the values of the acceleration parameters simulated).

Table 7 shows the mean value of the *a* parameter of the items administered. For all the selection rules, except S04, these values are clearly situated above the mean of the parameter in the bank, 1.3. Both PG and PP present a lower mean than PFI. The higher the value of the acceleration parameter, the more similar is the mean of the items administered to the mean of the bank. As in Study 1, content blocking leads to a reduction in the mean *a* value.

5.3. Discussion

In this second study, we find that the results of Study 1 are largely repeated. The PG and PP methods offer improvements in security and facilitate maintenance of the bank. As found in a condition in the first study, this can even be achieved with improvements in

Table 7. Mean *a* parameter of the presented items for different item selection rules and different acceleration parameters in Study 2 (the mean *a* value in the bank was 1.3)

	NCB	CB
PFI	1.56	1.52
PG^(−1)	1.54	1.51
PG^0	1.53	1.49
PG^1	1.50	1.46
PG^2	1.48	1.44
PP^(−1)	1.54	1.51
PP^0	1.53	1.50
PP^1	1.49	1.47
PP^2	1.47	1.45
S04	1.36	1.35

measurement accuracy, given that the alternative methods are less sensitive to restrictions in r^{max}.

For this bank, given that we consider the level of RMSE achieved with PFI to be satisfactory, it appears that the most suitable alternative will be to employ the method which, without loss of accuracy, offers the best results in the other variables. In this case, considering that in practice the test is administered without content balancing, the PP rule with an s parameter equal to 2 enables us to maintain the accuracy level while offering advantages in both security and bank maintenance.

6. Conclusions

Basing item selection, at the beginning of the CAT, mainly on chance, then gradually increasing the importance of information as the test progresses, improves the security of the bank and brings the distribution of the parameters of the items that will have to be replaced closer to the distribution of the items in the bank. In order to achieve this, we have described two item selection rules, the progressive method (Revuelta & Ponsoda, 1998) and the proportional method (Segall, 2004a). We have extended their original formulation, incorporating an acceleration parameter that defines the speed of the move from strictly random selection to selection based on information.

In Study 1, we showed how this could be achieved with very small losses of accuracy. We found similar results with a variety of conditions: two different test lengths, with and without content balancing and with and without restrictions on r^{max}. With the data shown, it seems that not even when the aim is to maximize accuracy is it advisable to use PFI. In practice, selection of the method to apply in a CAT will depend on the relative importance of the three components mentioned. The more important the accuracy, the lower the acceleration parameter will have to be. If security or maintenance of the bank is especially relevant, higher acceleration parameter values will be preferable or the S04 method should be used.

In the second study, we evaluated the impact of the different selection rules for the case of an operative bank. We largely replicated the results of Study 1, showing that our results were not limited to the characteristics of the simulated item banks. This study indicates that, for the uses of this bank, the most appropriate rule is PP with parameter s equal to 2.

Although the differences between PP and PG are slight, when the main goal is to improve security, the preferred method is PP. We suggest, as a tentative explanation, that the small differences found between the methods could be due to the fact that for the PG method there is no random component active for the last item, whereas for the PP method it continues to have a role for this item position. Both PG and PP present an overlap rate that is still above that which would be found with a uniform distribution of item-exposure rates. Likewise, for both methods the mean of the a parameter of the items administered is clearly above the mean of this parameter in the bank. If the priority is strict control of the exposure rates or maximum facility for replacement of withdrawn items, the use of the S04 method may be the most suitable alternative.

Acknowledgements

This research was partly supported by two grants from the Spanish Ministry of Education and Science (projects DGES-MEC BSO2002-01485 and SEJ2004-05872PSIC).

References

Barrada, J. R., Mazuela, P., & Olea, J. (2006). Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema, 18*, 156-159.

Barrada, J. R., Olea, J., & Ponsoda, V. (2007). Methods for restricting maximum exposure rate in computerized adaptative testing. *Methodology, 3*, 14-23.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2006). Estrategias de selección de ítems en un Test Adaptativo Informatizado para la evaluación de inglés escrito. [Item selection rules in a Computerized Adaptive Test for the assessment of written English]. *Psicothema, 18*, 828-834.

Chang, H. H. (2004). Understanding computerized adaptive testing - From Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 117-133). Thousand Oaks, CA: Sage Publications.

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, H. H., & Ying, Z. (1999). *a*-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

Chen, S. Y., & Ankenmann, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement, 41*, 149-174.

Chen, S. Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*, 241-255.

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*, 129-145.

Chen, S. Y., & Lei, P. W. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement, 29*, 204-217.

Dodd, B. G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14*, 355-366.

Eggen, T. J. H. M. (2001). *Overexposure and underexposure of items in computerized adaptive testing*. Measurement and Research Department Reports (2001-1). Arnhem, The Netherlands: CITO.

Hau, K. T., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38*, 249-266.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*, 359-375.

Leung, C. K., Chang, H. H., & Hau, K. T. (2003). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement, 63*, 257-270.

Li, Y. H., & Schafer, W. D. (2005). Increasing the homogeneity of CAT's item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. *Journal of Educational Measurement, 42*, 245-269.

Olea, J., Abad, F. J., Ponsoda, V., & Ximénez, M. C. (2004). Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: Diseño y comprobaciones psicométricas [A computerized adaptive test for the assessment of written English: Design and psychometric properties]. *Psicothema, 16*, 519-525.

Ponsoda, V., & Olea, J. (2003). Adaptive and tailored testing (including IRT and non IRT application). In R. Fernandez-Ballesteros (Ed.), *Encyclopaedia of psychological assessment* (pp. 9-13). London: Sage Publications.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Segall, D. O. (2004a). A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 29*, 439-460.

Segall, D. O. (2004b). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *The encyclopaedia of social measurement* (pp. 429–438). San Diego, CA: Academic Press.

Stocking, M. L., & Lewis, C. L. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163–182). Dordrecht, The Netherlands: Kluwer Academic.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63, 201–216.

van der Linden, W. J. (2003). Some alternatives to Sympson–Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 28, 249–265.

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, 42, 283–302.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–25). Dordrecht, The Netherlands: Kluwer Academic.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, 29, 273–291.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, 22, 203–226.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, 17, 17–27.

Received 16 December 2005; revised version received 11 July 2007

Capítulo 5.
Un método para la comparación de reglas de selección de items en tests adaptativos informatizados

Artículo enviado a revisión. La referencia del mismo es:

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (Enviado). *A method for the comparison of item selection rules in computerized adaptive testing*.

El texto que sigue corresponde al artículo tal y como ha sido enviado para su valoración.

Title

A method for the comparison of item selection rules in computerized adaptive testing

Abstract

In a typical study of the relative efficiency of two competing item selection rules in computerized adaptive testing, the common result is that they simultaneously differ in accuracy and security, making it difficult to reach a conclusion on which is the more appropriate rule. This study proposes a strategy to conduct a global comparison of two or more selection rules. A plot showing the performance of each selection rule for several maximum exposure rates is obtained and the whole plot is compared with other rule plots. The strategy has been applied in a simulation study for the comparison of 6 exposure control methods: point Fisher information, Fisher information weighted by likelihood, Kullback-Leibler weighted by likelihood, maximum information stratification method with blocking, progressive method and proportional method. Our results show that there is no optimal rule for any overlap value or RMSE. The fact that a rule, for a given level of overlap, has lower RMSE than another does not imply that this pattern holds for another overlap rate. A fair comparison of the rules requires extensive manipulation of the maximum exposure rates. The best methods were Kullback-Leibler weighted by likelihood, proportional method, and maximum information stratification method with blocking.

Two are, at least, two objectives to be maximized in a computerized adaptive test (CAT): the first is measurement accuracy; the second, item bank security. An item bank is considered more secure if the probability is low that an examinee is aware of the item content before a test is taken.

To evaluate the first objective, measurement accuracy, the variable usually employed has been the RMSE, computed according to Equation 1:

$$RMSE = \left(\sum_{g=1}^{r} \left(\hat{\theta}^g - \theta^g \right)^2 / r \right)^{1/2}, \quad (1)$$

where r is the number of examinees, θ^g is the trait level of the g-th examinee and $\hat{\theta}^g$ is the estimated trait level for that examinee.

The second goal, improvement in security, is related to the first. If an examinee receives an item that is known beforehand, for some CATS a correct response may be expected. As the probability of a correct response no longer depends on the examinee's trait level and on item parameters, test validity is compromised. The relevance of test security will vary between CATs. In some applications, such as personality measurement (e. g., Reise & Henson, 2000) or patient-reported outcomes (e. g., Cella, Gershon, Lai & Choi, 2007), examinees will probably be motivated for getting an assessment that is as accurate as possible, so security will become a minor issue. In the case of high-stake CATs, some examinees could try to inflate their score if they had the option (Chang, 2004; Davey & Nering, 2002).

The overlap rate, which has been used as an index of item bank security, is defined as the proportion of items that are shared, on average, by two randomly selected examinees (Way, 1998). The higher this is, the less secure the bank is. Chen, Ankenmann, and Spray (2003) have shown how the overlap rate is linearly related to the variance of the item exposure rates [$S^2_{P(A)}$]. When this variance approaches zero, the overlap rate approaches its minimum, Q/n, where Q is the test length and n is the item bank size. Throughout the paper, we will focus on fixed length CATs, so Q is constant for all the examinees. The overlap rate (T) can be calculated with Equation 2 (Chen et al., 2003):

$$T = \frac{n}{Q} S^2_{P(A)} + \frac{Q}{n}. \quad (2)$$

A wide range of item selection rules has been proposed. Some of them are focused, mainly, on increasing accuracy, while others seek to reduce security risks. In general, a trade-off between accuracy and security has been found: increments in one variable mean reductions in the other. It is assumed that this trade-off holds both within- and between-rules (Chang & Ansley, 2003; Stocking & Lewis, 2000). For any given rule, it is understood that any manipulation which is effective in reducing the overlap rate (as, for instance, restrictions of the maximum exposure rate – see below), will yield increments in RMSE. For any two different rules, it is expected that the rule with the lower RMSE will necessarily be the one with greater overlap rate.

An example may help to explain this better. Let us imagine that we want to compare X and Y item selection rules. For the X rule, RMSE is .25 and overlap rate is .18. For the Y rule, the resulting RMSE is .27 and overlap rate is .12. This would be an example of the abovementioned trade-off. Results of this kind are usually described with sentences like this: "The Y rule is the most convenient, as it allows security improvements with a small impact in accuracy". From our point of view, in conditions where there is a trade-off between accuracy and security, it is not possible to draw any conclusion. The X rule could be modified by incorporating, for instance, restrictions on the maximum exposure rate, which may yield an overlap rate equal to that obtained by the Y rule. In that case, we do not know if the RMSE found with the X rule would be greater, lower or equal to the RMSE of the Y rule. Any conclusion drawn from studies where no item selection rule dominates over the other (it is better in terms of one variable and equal or better in terms of the other) should be considered with caution.

Our aim is to present a method that allows a better comparison between different item selection rules. The paper is structured as follows. First, we will show a comparison procedure that will allow us to establish, for a given level of accuracy or security, which selection rule is to be preferred. Second, a method for restricting the maximum exposure rate will be presented. Third, some of the rules that have been proposed until now will be described and, fourth, we will illustrate the proposed method with a simulation study where 6 item selection rules are compared.

A method for the comparison of item selection rules

As commented above, two item selection rules should share the same value in one of the variables of interest (accuracy or security) for a fair comparison of their efficiency. When this happens, a safer conclusion on their relative quality can be drawn by comparing their

performance in the other variable. The probability that two selection rules show, without any additional manipulation, the same value in one of the variables is, of course, very small. This probability is even smaller if the number of rules to be compared is more than two and all of them should share the same value for one variable.

The use of methods that restrict the maximum exposure rate of the items (r^{max}) has been the most common solution to this problem, as they are effective in reducing the overlap rate of the rule with lower bank security. A r^{max} value that can provide a similar overlap rate for both selection rules is obtained and then the measurement accuracy obtained with each of the rules is compared. Chang and Ying (1999) provide one example of this approach.

This method presents, at least, two limitations. First, as the r^{max} values used for the comparison are tentatively established, the overlap rates for all the rules usually show some differences. Second, although for a given common overlap rate the RMSE obtained with one selection rule could be lower than the RMSE of another rule, it should not be taken for granted that this pattern of results will hold for any other security level. It could be possible that the item selection rule to be preferred varies according to the security control desired.

An alternative strategy for comparing as many selection rules as needed for the whole range of possible values of accuracy and security can be achieved by manipulating r^{max}. Our proposal is to manipulate the r^{max}, for each item selection rule, in V different values, ranging from $r_1^{\max}$ equal to Q/n, which is the minimum possible value for r^{max}, to $r_V^{\max}$ equal to 1, which is equivalent to not applying any restriction on the maximum exposure rate. This idea mimics the method used by Barrada, Olea and Abad (2008) for the comparison between rotating items banks and the restriction on maximum exposure rates in a master bank.

With this strategy, we obtain tables of results with a structure as shown in Table 1: we find RMSE and overlap rate for V differents conditions, starting with maximum item exposure control and finishing with no item exposure control. We have one independent variable (r^{max}) and two dependent variables (RMSE and overlap). With these data, it possible to obtain the curves that relate r^{max} with RMSE and r^{max} with the overlap rate. Also, with this information it is possible to plot the graph that relates the overlap rate with RMSE.

Table 1
Relation between the indicator variable *v* and the corresponding r^{max}, overlap rate and RMSE

v	r^{max}	Overlap rate	RMSE
1	$r_1 = Q/n$	T_1	$RMSE_1$
2	r_2	T_2	$RMSE_2$
...	...	...	...
V	$r_V = 1$	T_V	$RMSE_V$

As many tables as item selection rules to be compared are generated. In this way, we can plot a curve for each item selection rule. Thus, it is possible to make the comparisons we wanted: holding RMSE (or overlap) constant we can check the differences in overlap (or RMSE). Imagine that the X-axis represents the overlap rate and the Y-axis corresponds to the RMSE. If the curve of an item selection rule is always below the curve for another item selection rule, the former should be preferred, as for any value of in one variable it offers better results in the other variable. If two curves cross at some point, this means that the optimal selection rule depends on the degree of security (or accuracy) that is desired and no rule dominates the other over the whole range of possible values.

Following Barrada, Olea and Abad (2008), the different r^{max} values are defined by means of Equation 3:

$$r_v^{\max} = \frac{Q}{n} + \frac{\left(1 - Q/n\right)\sum_{f=1}^{v}(f-1)^2}{\sum_{f=1}^{V}(f-1)^2}, \tag{3}$$

where *v* is used for defining the position in the *V* different r^{max} values ($r_1^{\max}$ the minimun and $r_V^{\max}$ the maximum).

This equation leads to unequally spaced values of $r_v^{\max}$, a characteristic that is desirable, given the usual form of the curves relating overlap rate with RMSE (Barrada, Olea & Abad, 2008).

As an important part of this comparison method relies on the control of maximum exposure rates, we present now one of the methods proposed for doing so.

Restriction of maximum exposure rate

The most common approach to improving item bank security is to reduce r^{max} (van der Linden, 2003). The methods aimed at restricting maximum exposure rate eliminate the problem of item overexposure and reduce the overlap rate, although with an increase in RMSE. Each method establishes the probability for an item being eligible - $P(E_i)$- for administration. This probability will be lower for items with higher exposure rates (when no restriction on r^{max} is applied) than for underexposed items. The methods differ in how the $P(E_i)$ values are computed and in their range. When the Sympson-Hetter method (Sympson & Hetter, 1985), the restricted method (Revuelta & Ponsoda, 1998), and the item-eligibility method (van der Linden & Veldkamp, 2004) are compared, the proposal put forward by van der Linden and Veldkamp is the one that seems to be preferable (Barrada, Abad & Veldkamp, 2009).

In the item-eligibility method, the calculation of the $P(E_i)$ parameters for the *(m+1)*-th examinee is made according to Equation 4:

$$P^{(m+1)}(E_i) = \begin{cases} 1 & if \quad P^{(1..m)}(A_i)/P^{(m)}(E_i) \le r^{\max} \\ r^{\max} P^{(m)}(E_i)/P^{(1..m)}(A_i) & if \quad P^{(1..m)}(A_i)/P^{(m)}(E_i) > r^{\max} \end{cases}, \tag{4}$$

where $P^{(1..m)}(A_i)$ is the probability of the administration (exposure rate) of the *i*-th item computed

from the responses from the first to the *m*-th examinee. Further details of this method can be found in van der Linden and Veldkamp (2004, 2007).

We have presented the general framework that will allow us to compare different item selection rules. The next step is to present some of these rules, the ones that will be compared by means of the proposed method in a simulation study.

Item selection rules

Our selection of specific item rules is based on their relevance to our questions, as well as the amount of research available on their performance. The descriptions will not be exhaustive (more information can be obtained from the references).

Point Fisher information (PFI)

The item selection rule most commonly employed in CATs for the selection of the *q*-th item, *q* being the indicator of the item's position on the test, is the selection of the item with

maximum Fisher information for the estimated trait level (PFI rule; Lord, 1980). The selected item *j* is given by:

$$j = \arg\max_{i \in B_q} I_i(\hat{\theta}) \quad (5)$$

where $I_i(\hat{\theta})$ is the Fisher information of the item *i* for the estimated trait level and B_q is the subset of items belonging to the item bank that can be presented to the examinee in the *q*-th position in the test. If no restriction is active, B_q consists of those items not presented to that examinee in the *q*-1 previous items. If the item-eligibility method is used, B_q is the intersection between the non-presented items and those items marked as eligible.

Fisher information weighted by likelihood (FI-L)

The PFI rule has several limitations. On the one hand, it does not take into account the values of the information function for trait levels different from the estimated trait level. On the other, the likelihood function $[L(\theta)]$ is merely used to locate its maximum, playing no role at all its shape, which can vary from being mainly flat, as at the beginning of the test, to more peaked, as the test goes on. Additionally, it does not take into account the possibility of various local maxima in the likelihood function (Samejima, 1977). Veerkamp and Berger (1997) proposed a more exhaustive use of both functions with the item selection rule called Fisher information weighted by likelihood (FI-L), described in Equation 6:

$$j = \arg\max_{i \in B_q} \int_{-\infty}^{\infty} I_i(\theta) L(\theta) d(\theta). \quad (6)$$

The entire trait level range affects the FI-L rule. This allows for greater accuracy of this rule when compared with PFI (Veerkamp & Berger, 1997), especially for low trait levels (Chen, Ankenmann & Chang, 2000), although this is achieved with an increment in the overlap rate (Chen & Ankenmann, 2004).

Kullback-Leibler function weighted by likelihood (KL-L)

The Kullback-Leibler (KL) information function evaluates the item discrimination capacity between any possible pairs of trait levels. This means that KL is a global information measure (Chang & Ying, 1996). The Kullback-Leibler function weighted by likelihood (KL-L) rule is defined in Equation 7:

$$j = \arg\max_{i \in B_q} \int_{-\infty}^{\infty} KL_i(\theta \,||\, \hat{\theta}) L(\theta) d(\theta), \quad (7)$$

where $KL_i(\theta \,||\, \hat{\theta})$ is calculated as follows:

$$KL_i(\theta \| \hat{\theta}) = P_i(\hat{\theta}) \ln\left[\frac{P_i(\hat{\theta})}{P_i(\theta)}\right] + \left[1 - P_i(\hat{\theta})\right] \ln\left[\frac{1 - P_i(\hat{\theta})}{1 - P_i(\theta)}\right]. \quad (8)$$

When compared with the PFI, KL-L offers lower RMSE (Chen et al., 2000), although with a greater overlap rate (Chen & Ankenmann, 2004).When compared with the FI-L, KL-L reduces RMSE and increases the overlap rate (Barrada, Olea, Ponsoda & Abad, 2009).

Maximum information stratification method with blocking (MIS-B)

The logic of stratified methods, first proposed by Chang and Ying (1999), is to administer low-informative items at the beginning of the test, and to increase the administration of more highly informative items as the test goes on. In all the stratified methods, those items belonging to B_q are determined according to their position in the test length. We will follow the formulation of the method proposed by Barrada, Mazuela and Olea (2006), as it outperforms the original in both security and accuracy.

In the 3-parameters logistic model, the maximum Fisher information of item *i* ($I_i^{\max}$) is equal to (Hambleton & Swaminathan, 1985):

$$I_i^{\max} = \frac{1.7^2 a_i^2}{8\left(1 - c_i^2\right)}\left[1 - 20c_i - 8c_i^2 + \left(1 + 8c_i\right)^{3/2}\right], \quad (9)$$

where a_i is the discrimination parameter and c_i is the pseudo-guessing parameter for item *i*.

The trait level at which this maximum information is achieved ($\theta_i^{\max}$) can be calculated according to Equation 10 (Hambleton & Swaminathan, 1985):

$$\theta_i^{\max} = b_i + \frac{\ln\left[1 + \left(1 + 8c_i\right)^{1/2}\right] - \ln(2)}{1.7a_i}, \quad (10)$$

where b_i is the location parameter of item *i*.

In the MIS-B method, prior to any item being administered, the item bank is stratified. First, the *n* items of the bank are ordered in increasing order according to their $\theta_i^{\max}$ values. The first *S* items (*S* being the number of strata into which the item bank will be divided) are rearranged, ordering them in an ascending order according to their $I_i^{\max}$ value. The first item of this *S* item set will be assigned to the first stratum, the second to the second... and the *S*-th item to the *S*-th stratum. This process is repeated for the n/S blocks of size *S* that can be obtained.

In a CAT of length *Q*, during the first Q/S items of the test administration, the B_q item set will be formed by the n/S items of the first stratum, the n/S items of the second stratum will compose B_q for the next Q/S items of the test, and so on. As the test goes on, the mean $I_i^{\max}$ of the items belonging to B_q increases, leaving the items with high *a* parameter and low *c* parameter ready for use at the end of the test. Stratifying by taking into account $\theta_i^{\max}$ (blocking it) makes the distribution of $\theta_i^{\max}$ as similar as possible between strata. Once the B_q set is defined for each item, the selection will be made according to Equation 11:

$$j = \arg\min_{i \in B_q} \left| \hat{\theta} - \theta_i^{\max} \right|. \tag{11}$$

The stratified methods, compared with PFI, improve the security of the item bank, leading to an overlap rate near the minimum possible overlap rate (e. g., Chang, Qian and Ying, 2001; Cheng, Chang & Yi, 2007), while decreasing accuracy (Chang & Ying, 1999).

Progressive method (PG)

Revuelta and Ponsoda (1998) proposed the progressive method (PG). This method selects the item for which the sum of a random component and the Fisher information is highest. At the beginning of the test, when the trait estimation error is high, the weight of the random component (W_q) is most important. The weight of the Fisher information increases as the number of items administered increases. The PG method can be described as follows:

$$j = \arg\max_{i \in B_q} \left[(1 - W_q) R_i + W_q I_i(\hat{\theta}) \right], \tag{12}$$

where the weight W_q is the contribution of item information to the selection criterion and R_i is a random number belonging to the interval [0, $\max_{i \in B_q} I_i(\hat{\theta})$].

Barrada, Olea, Ponsoda and Abad (2008) have proposed the following equation to relate W_q to *q*:

$$W_q = \begin{cases} 0 & \text{if} \quad q = 1 \\ \dfrac{\sum_{f=1}^{q} (f-1)^r}{\sum_{f=1}^{Q} (f-1)^r} & \text{if} \quad q \neq 1 \end{cases}. \tag{13}$$

The *t* parameter marks the speed at which the weight of the random component is reduced and, thus, the speed at which the importance of item information increases. This parameter defines the improvement in bank security and accuracy reduction, in comparison with PFI. With a *t* equal to 1, marked improvements in security are obtained with hardly any impact on accuracy.

Proportional method (PP)

All the methods presented so far are deterministic, in the sense that the item to be selected maximizes (or minimizes, in the case of the stratified methods) the selection function. Segall (2004) has proposed a stochastic method, where the selection function value is not used to order the items and to select the first item, but to calculate the probability of selecting each item. Hence, no item would have a probability of being administered equal to 0. In this method, which will be called proportional (PP; Barrada, Olea, Ponsoda & Abad, 2008; Segall, 2004), the probability of selecting the items is given by Equation 14:

$$P(S_i) = \frac{(1-z_i)I_i(\hat{\theta})^{H_q}}{\sum_{i=1}^{n}(1-z_i)I_i(\hat{\theta})^{H_q}}, \tag{14}$$

where z_i indicates whether the item belongs (1) or not (0) to B_q. Once the probabilities of each item being selected are obtained, a cumulative distribution of probabilities is formed. Then, a random number drawn from the uniform interval (0,1) is used to identify the item to be selected.

When H_q is 0, selection is random. The higher H_q, the higher the probability of selection of the item with maximum Fisher information will be, making the selection by the PP and PFI methods more similar.

Barrada, Olea, Ponsoda and Abad (2008) have proposed defining H_q according to Equation 15, which, as can be seen, clearly resembles the equation that defines the values of W_q for the PG method:

$$H_q = \begin{cases} 0 & if \quad q = 1 \\ \dfrac{Q\sum_{f=1}^{q}(f-1)^s}{\sum_{f=1}^{Q}(f-1)^s} & if \quad q \neq 1 \end{cases}. \tag{15}$$

According to this function, the test starts with random selection and, for common CAT lengths, the selection of the last item will be similar to selection with PFI. The *s* parameter has the role the *t* parameter has in the PG rule: it defines the speed at which the method reduces random selection of items.

The 6 rules described are, in our view, an acceptable sample of the range of the available selection rules for CAT: PFI is the current standard in item selection; while FI-L and KL-L illustrate the pole of maximum accuracy with minimum security, MIS-B is at the opposite extreme, with a low overlap rate with accompanying increments in RMSE; finally, PG and PP are methods that obtain satisfactory results in accuracy, comparable to those obtained with PFI, with improvements in security. More importantly, these 6 item selection rules described will allow us to show the use of our proposed comparison procedure.

Simulation study

Method

Wingersky and Lord (1984) and Chang, Qian and Ying (2001) have pointed that, in practice, the *a* and *b* parameters of the items are usually correlated. The performance of different item selection rules can vary depending on whether the item banks employed have correlated parameters or not (Barrada, Olea et al., 2009). Thus, two kinds of item banks were generated, one with uncorrelated *a* and *b* parameters and the other with correlated parameters ($r_{ab} = .5$). A total of 10 banks of 500 items each were obtained for the correlated and uncorrelated bank types. The parameter distributions were: $a \sim N(1.2, .25)$, $b \sim N(0, 1)$ and $c \sim N(.25, .02)$. For each item bank 5,000 examinees were generated randomly, with trait levels extracted from a distribution N(0, 1). Two different test lengths, 20 and 40 items, were used. The initial trait level, $\hat{\theta}_0$, was selected randomly from the uniform interval (-.5, .5). Dodd's (1990) procedure was applied for the trait level estimation until each examinee obtained correct and incorrect responses: when all the responses were correct, $\hat{\theta}$ was increased by $(b_{max} - \hat{\theta})/2$; if all the responses were incorrect, $\hat{\theta}$ was reduced by $(\hat{\theta} - b_{min})/2$. Once the constant pattern was broken or the test was finished, maximum-likelihood estimation was applied, with the restriction that $\hat{\theta}$ had to be in the interval [-4, 4].

The likelihood function cannot be computed when no item has been administered. Because of this, for the selection of only the first item with the FI-L and KL-L rules, two fictitious items were added to the response vector to obtain the likelihood function, one correct and one incorrect, both with the same parameters: a=.5, b= $\hat{\theta}_0$ and c=0 (Barrada, Olea et al., 2009).

In the MIS-B rule, the bank was stratified into 5 strata, all of them having the same number of items. The number of items extracted from each stratum was held constant. In the PG method, the *t* parameter was set equal to 1. The same value was given for the *s* parameter in the PP method.

For the restriction of the maximum exposure rates, the item-eligibility method was used (van der Linden & Veldkamp, 2004).

The variables for comparing the different methods were RMSE and the overlap rate, as defined in Equations 1 and 2. The strategy of simulating several different r^{max} values for each item selection rule was applied. Barrada, Olea and Abad (2008) showed that 10 values for r^{max} are enough to a plot the desired graphs correctly, so we fixed *V* (Equation 3) at 10.

Results

The relation between the overlap rate and r^{max} can be seen in Figure 1. Results for each dot in the plot were based on 50,000 examinees (10 banks * 5,000 simulees). The results maintain the same pattern independently of the number of items administered or the correlation between parameters. When no restriction on maximum rate is applied (i.e., r^{max} equal to 1) the expected results were found: the KL-L rule produced the highest overlap rates, followed by FI-L and, after these, PFI. Higher security is achieved with the PG and PP rules. When the test length is 20 items, PP is more secure than PG; however, with 40-item tests both rules offer the same overlap rate. The rule that offers the highest security level is MIS-B, as its overlap rate, even when r^{max} is 1, is very close to the minimum possible value.

The effect of restricting r^{max} is not the same for all the selection rules. While for the rules with greater overlap when r^{max} is 1, small changes in r^{max} reduce overlap; for the rules with better exposure control, a greater reduction in r^{max} is needed to improve security. In other words, the rules with greater overexposure problems are more sensitive to changes in r^{max}.

The effects of reducing r^{max} on the overlap rate are more evident with low r^{max}: the effect of changing r^{max} from 1 to .9 is smaller than the effect of changing from.2 to .1. In any case, when we impose the minimum possible r^{max}, we obtain minimum overlap.

Figure 2 shows the relation between r^{max} and RMSE. As expected, increasing test length improves accuracy. The correlation among parameters increases RMSE, as in these banks the information available around the average trait levels is lower than in banks with uncorrelated parameters. Increasing test length reduces the differences in accuracy between rules.

When there is no restriction on r^{max}, the selection rule that offers greater accuracy is KL-L, followed by FI-L, although in 40-item tests the difference is negligible. The rule with greater measurement error is MIS-B. For a length of 20 items and correlated parameters, PFI obtains higher RMSE than PG and PP, while for the rest of the conditions, PFI is more accurate.

The different selection rules allow important restrictions on r^{max} without translating this into increments in RMSE. Fixing, for instance, r^{max} to .3 does not imply any noticeable losses in accuracy. As r^{max} approaches its minimum possible value, the speed with which RMSE increases is accelerated.

Figure 3 depicts the relation between the overlap rate and RMSE. These plots are, in our view, the most relevant for deciding which selection rule to choose for a CAT. As expected, there is a within-rule trade-off between accuracy and security. However, important improvements in security can be achieved with negligible reductions in accuracy. The main comparison is between rules. The rule to be selected will depend on the required accuracy level and on the acceptable security risk. Let us consider first the results for a test length of 20 items. If our CAT tolerates overlap rates over .11, in the case of uncorrelated banks, or over .08 when r_{ab} is equal to .5, the rule to be used is KL-L, as for these overlap rates it provides the highest accuracy levels. If a more strict control of bank security is needed, the most convenient rule is PP. Looking at the same data from the perspective of accuracy, if we want RMSE values between .27 (minimum possible value) and .30, for the condition of uncorrelated banks, or between .28 and .32 in the correlated case, the rule to use is KL-L. If a reduction in accuracy in order to increase security seems appropriate, the best alternative is PP. None of the other four rules tested would be selected for any level of security or accuracy.

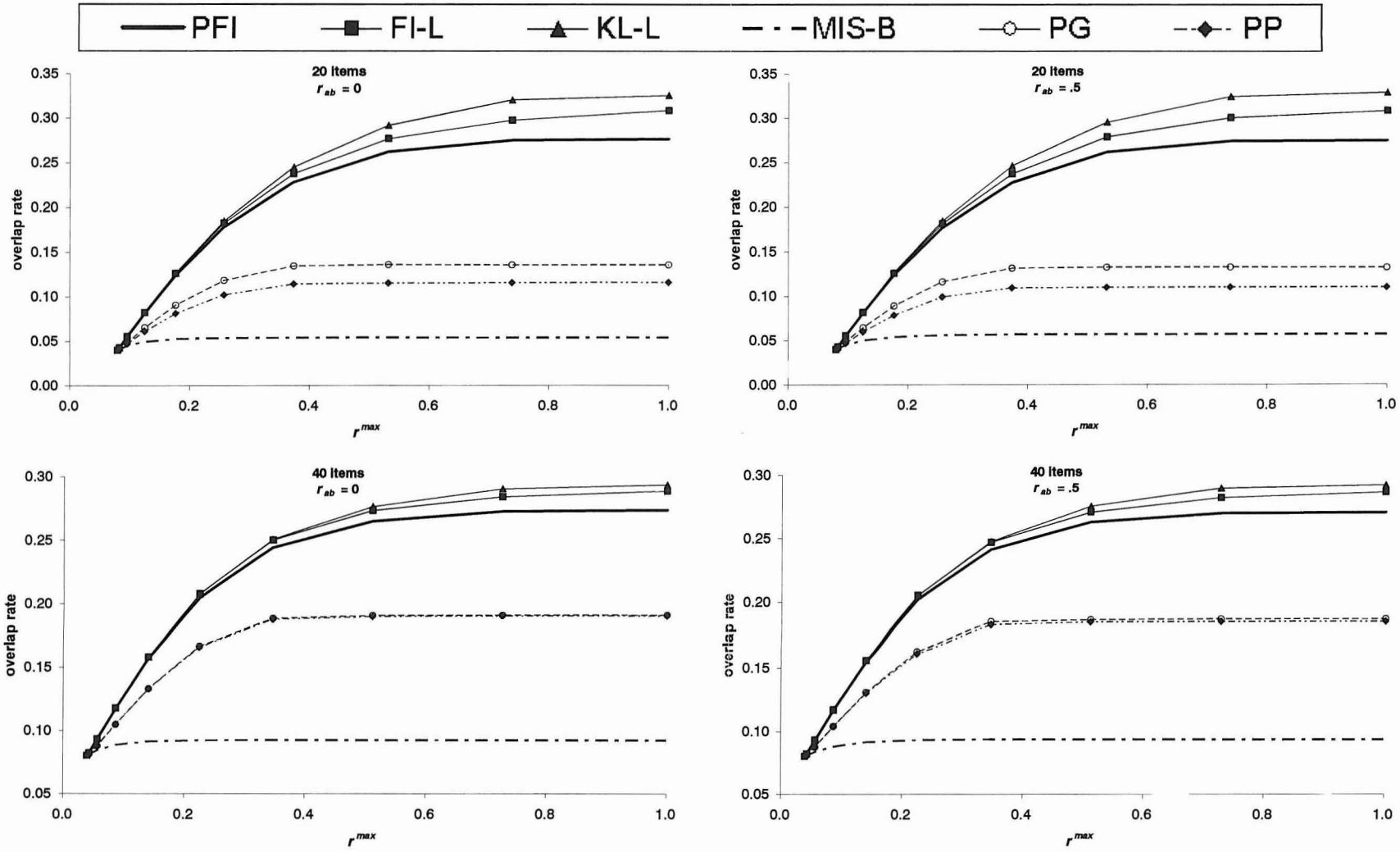


Figure 1. Relation between r^{max} and overlap rate.

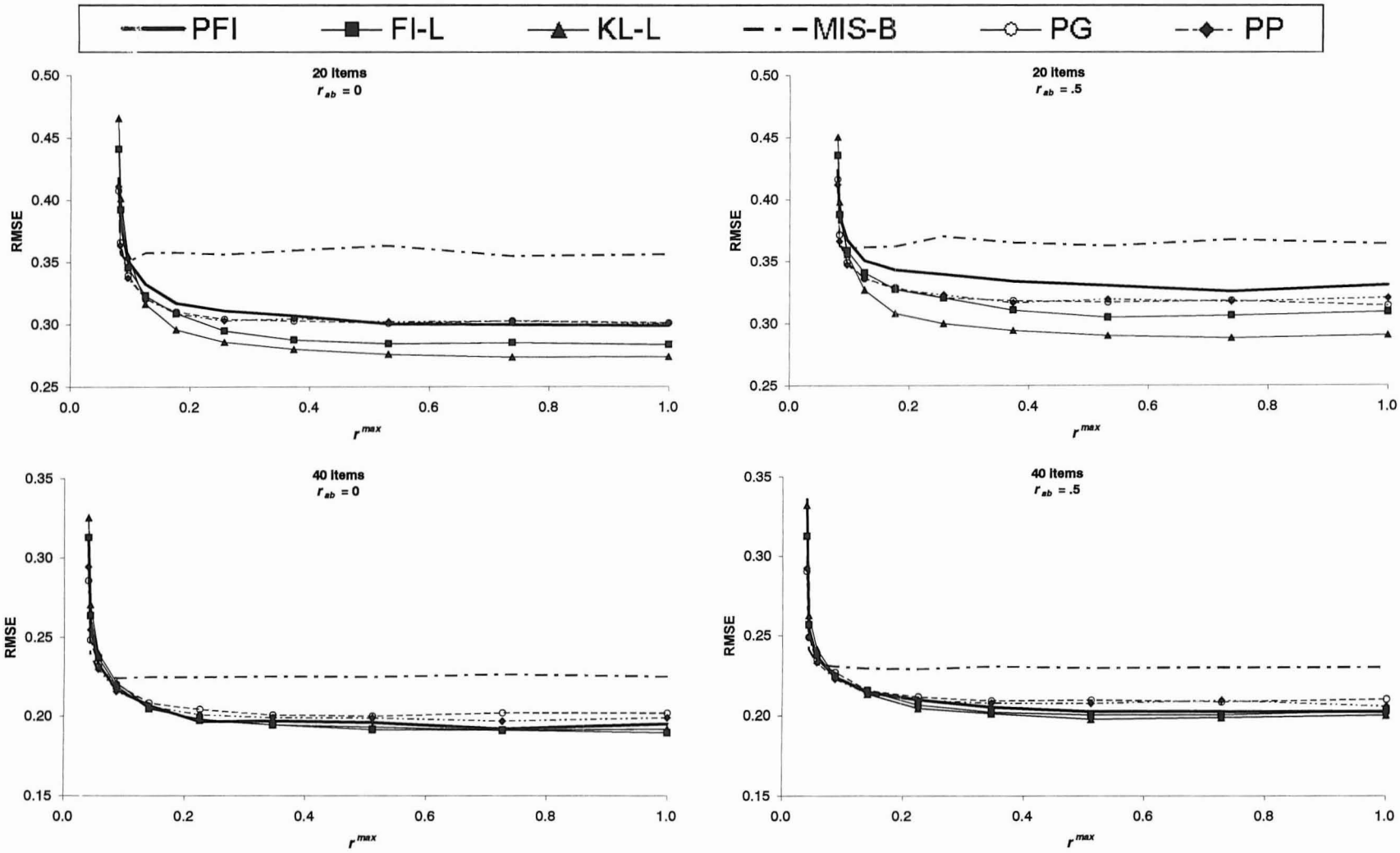


Figure 2. Relation between r^{max} and RMSE.

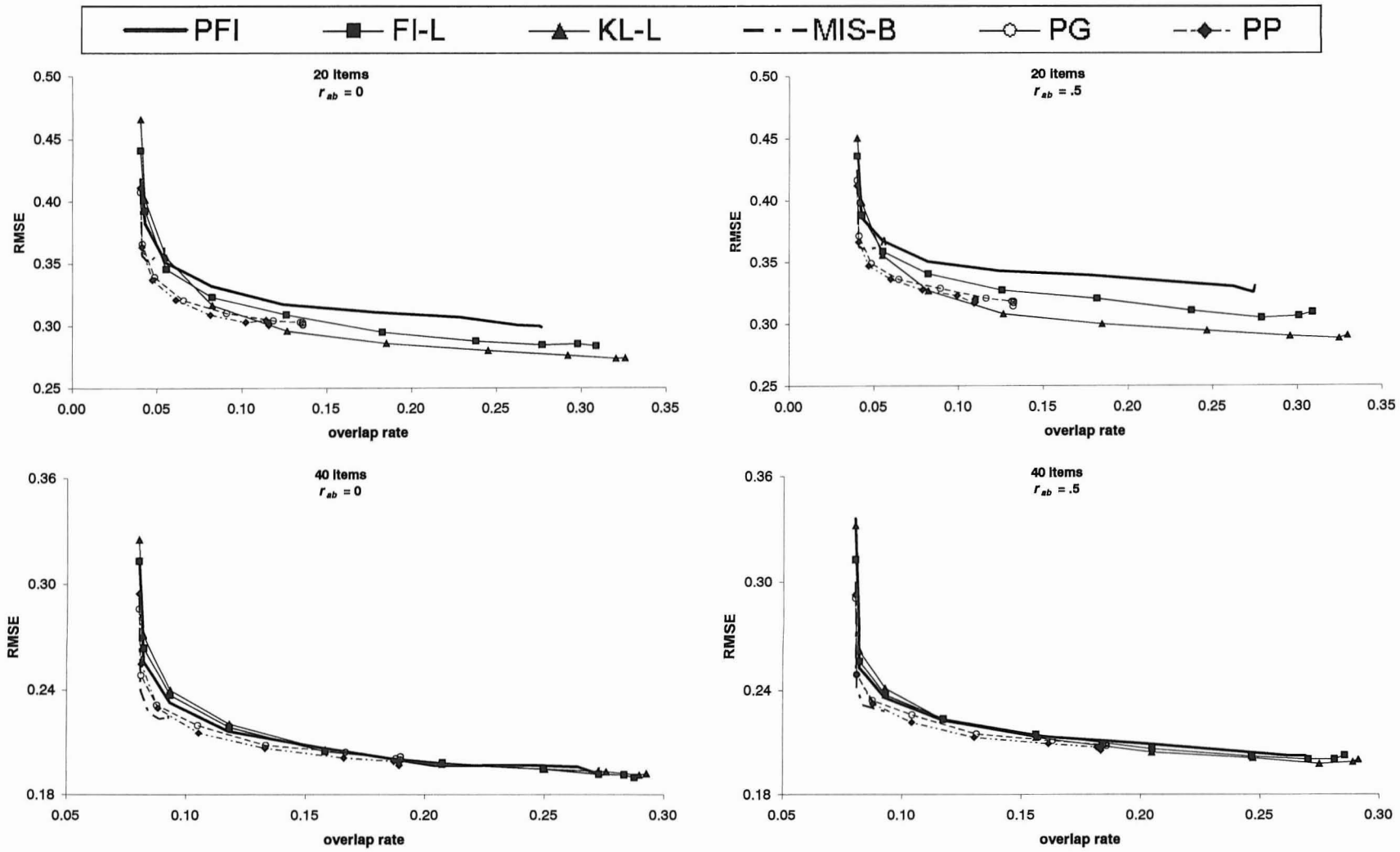


Figure 3. Relation between overlap rate and RMSE.

The differences between selection rules when the test length is 40 items are much smaller. When maximizing accuracy is the main objective or when high overlap rates can be tolerated, the most convenient rules are KL-L, FI-L and PFI. When a greater exposure control is desired and we can tolerate just a small increment in measurement error, the rule to be used is PP. If we want the highest possible item bank security, given these conditions of test length and bank size., the rule to choose is MIS-B.

Discussion and conclusions

Several item selection rules are available for CAT. Some of them focus on measurement accuracy, like PFI, FI-L and KL-L, while others are focused on item bank security, like MIS-B, PG and PP. The comparison of the results provided by different selection rules is not an easy task, as rules offering better accuracy are usually lower on security indicators. The proposed strategy enables an improved comparison of item selection rules, as rules can be compared in one indicator (accuracy or security) while holding the other constant. Two main features of the proposed strategy are, first, that it can be easily applied to more than two selection rules and, second, that it compares the rule's global performance, rather than just its efficiency for a particular pair of accuracy-security values.

The strategy was applied for the comparison of 6 selection rules and provided these main results:

(a) The item selection rule most commonly employed, PFI, is never the best alternative. At most, we could recommend its use for a test of 40 items with uncorrelated parameters when poor exposure control can be tolerated or test security is a minor issue, although in this case its performance is equivalent to that provided by KL-L and FI-L. A possible reason to prefer PFI in these conditions could be it is lower computational complexity, although we consider that, with modern computers, differences will be negligible in terms of CPU time.

(b) The FI-L rule seems, also, to be outperformed almost continuously by KL-L. The PP rule is always a slightly better alternative than PG. So, it seems that three out of the 6 rules (PFI, FI-L and PG) could be discarded.

(c) There is no an optimal rule for any value of overlap or RMSE. The fact that a rule, for a given level of overlap, has lower RMSE than another does not imply that this pattern has to hold for another overlap rate. So, a fair comparison of the rules requires an extensive manipulation of the maximum exposure rates, as the proposed strategy does, in order to obtain more than just one pair of accuracy-security indicators and to

enable the comparison of the global efficiency curve of one selection rule against the others. Studies lacking such an extensive manipulation should be considered with caution.

(d) The point at which PP becomes preferable to KL-L depends on the kind of item bank employed and on the test length. The MIS-B rule is a viable option only for a test length of 40 items.

Some limitations of the method proposed should be noted. Although we have been using overlap rate as the only variable for assessing test security, several other variables have been proposed, mainly χ^2 (Chang & Ying, 1999), number of items unused from the bank and maximum exposure rate. In plots such as Figure 3, the overlap rate is held constant, but the other indicators of security could differ between rules; thus we could be taking as equal security conditions that are, in fact, unequal. For performing the comparisons, a single measure of security is needed. In this way, readable plots can be drawn. We have offered an imperfect but interpretable solution. Also, we believe that the other variables for measuring test security are redundant or more limited.

(a) χ^2 is (Chang & Ying, 1999):

$$\chi^2 = \frac{\sum_{i=1}^{n}\left[P(A_i) - Q/n\right]^2}{Q/n}. \tag{16}$$

χ^2 is a measure of departure from uniform usage of items. With some substitutions, it can be shown that χ^2 is equal to $(nT - Q)$. For item banks and tests of the same length, χ^2 is a linear transformation of *T*, so the ordering of rules will not change.

(b) The overlap rate takes into account the whole distribution of item exposure rates to produce a single number, which is easy to interpret. Maximum exposure rate and number of items are restricted to just one of the extremes of this distribution.

Because of this, we believe that the overlap rate is the measure to be used in these conditions, when only one value is desired. Along this line, Yi, Zhang and Chang (2008) have shown that, holding constant the maximum exposure rate, the item selection rule with lower overlap rate, the alpha-stratified method, could better resist an environment of item bank disclosure, when compared with PFI, thus indicating that the overlap rate is a valid measure of test security.

The results of this study provide general guidance on choosing an item selection rule. The simulation conditions do not exhaust all the variables that could be relevant. We

have not considered, for instance, different item bank sizes or other parameter distributions. So, for the final decision on an optimal rule for a specific item bank and goals, our advice is to conduct an ad-hoc simulation study including the KL-L, PP and MIS-B rules. Having decided beforehand the level of accuracy or security we are looking for, a plot such as that shown in Figure 3 would allow us to decide which selection rule better fits our needs. Plots like Figure 1 or Figure 2 would help us to establish the r^{max} value needed to obtain our target values.

References

Barrada, J. R., Abad, F. J., & Veldkamp, B. P. (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema, 21*, 313-320.

Barrada. J. R., Mazuela, P., & Olea, J. (2006). Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema, 18*, 156-159.

Barrada, J. R., Olea, J., & Abad, F. J. (2008). Rotating item banks versus restriction of maximum exposure rates in computerized adaptive testing. *The Spanish Journal of Psychology, 11*, 618-625.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness to the Fisher information for improving item exposure control in CATs. *British Journal of Mathematical and Statistical Psychology, 61*, 493-513..

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2009). Item selection rules in computerized adaptive testing: Accuracy and security. *Methodology, 5*, 7-17.

Cella, D., Gershon, R., Lai, J., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research, 16*, 133-141.

Chang, H. H.. (2004). Understanding computerized adaptive testing – From Robbins-Monro to Lord and beyond. In David Kaplan (Ed.) *The SAGE handbook of quantitative methodology for the social sciences* (pp. 117-133). Thousand Oaks, CA: Sage Publications.

Chang, H. H., Qian, J., & Ying, Z. (2001). a-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement, 25*, 333-341.

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, H. H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

Chang, S. W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40*, 71-103.

Chen, S. Y., & Ankenmann, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of Computerized Adaptive Testing. *Journal of Educational Measurement, 41*, 149-174.

Chen, S. Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*, 241-255.

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*, 129-145.

Cheng, Y., Chang, H. H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement, 31*, 467-482.

Davey, T., & Nering, N. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward, (Eds). *Computer-based testing: Building the foundation for future assesments* (pp. 165-191). Mahwah, NJ: Lawrence Erlbaum.

Dodd, B. G. (1990) The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14*, 355-366.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Hingham, MA: Kluwer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Reise, S. P., & Henson, J. M. (2000). *Computerization and adaptive administration of the NEO-PI–R. Assessment, 7*, 347–364.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 233-247.

Segall, D. O. (2004). A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 29*, 439-460.

Stocking, M. L., & Lewis, C. L. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice* (pp. 163-182). Dordrecht, The Netherlands: Kluwer Academic.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden. W. J. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 28*, 249-265.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273-291.

van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics 32,* 398-418.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203-226.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17-27.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*, 347-364.

Yi, Q., Zhang, J., & Chang, H. H. (2008). Severity of organized item theft in computerized adaptive testing: A simulation study. *Applied Psychological Measurement, 32,* 543-558.

Capítulo 6.

Filtrado del banco de items en tests adaptativos informatizados. ¿Qué hace más segura a una regla de selección de items?

Artículo enviado a revisión. La referencia del mismo es:

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (Enviado). *Item bank disclosure in computerized adaptive testing: What makes an item selection rule safer?*

El texto que sigue corresponde al artículo tal y como ha sido enviado para su valoración.

Title

Item bank disclosure in computerized adaptive testing: What makes an item selection rule safer?

Abstract

A computerized adaptive test is considered more secure the lower the overestimation of the examinee's trait level due to item pre-knowledge. The common measures of test security have been the overlap rate between examinees and the distribution of item exposure rates. We explain that lower overlap rates or less homogeneous distributions of usage of the items may not lead to safer CATs. Instead of these variables, we show that the probability of item pre-knowledge of the first items administered and the overlap rate for high trait levels are better variables for assessing test security. If low values are present for these two variables, there are many different routes to obtain an estimated high trait level and, thus harder for an examinee with item pre-knowledge to incorporate to one of these routes. This is illustrated in three different studies where item bank disclosure is simulated. In these studies we compare the point Fisher information, the progressive method and the alpha-stratified selection rules. The alpha-stratified method, the option leading to lower overlap rate and more homogeneous item exposure distribution when there is no bank disclosure, is not the selection rule offering higher test security.

The risk of examinees receiving inflated trait estimates due to previous item knowledge is one danger in computerized adaptive testing (CAT; Chang, 2004). CAT allows continuous testing with an item bank that is static over time. This characteristic of CAT makes it possible for future examinees to obtain information from previous examinees about the items they received. Test security is, thus, a major concern in CATs (Davey & Nering, 2002).

We will consider a CAT to be more secure the lower the benefit in trait level estimation that an examinee could obtain due to item pre-knowledge. Two main variables have been used as indicators of test security (Chang & Zhang, 2002). First, the pair-wise overlap rate, which is defined as the proportion of items that, on average, two examinees share (Way, 1998). It has been assumed that the greater the overlap rate the greater the trait level overestimation due to item bank disclosure. Second, the distribution of item exposure rates, under the assumption that CATs with more homogeneous rates will be more robust to item leakage (Chen, Ankenmann & Spray, 2003).

Limitation of security evaluation in CATs

Usually, the overlap rate and the distribution of the item exposure rates are obtained by means of studies where no examinee has item pre-knowledge. In these studies, the probability of a correct response is determined by a typical IRT model, where no parameter marking the presence or absence of pre-knowledge of each item is included. A fewer number of studies have simulated item sharing (McLeod & Lewis, 1999; McLeod, Lewis, & Thissen, 2003; Mills & Steffen, 2000; Segall, 2002; Stocking, Ward, & Potenza, 1998; Yi, Zhang, & Chang, 2008), although none of them have questioned the idea that lower overlaps rate or more balanced usage of items lead to higher security.

In our point of view, when evaluating CAT security without including bank disclosure, we could be missing some important aspects that are present when there is item pre-knowledge. We will illustrate this with some examples where we will show that a lower overlap rate may not lead to higher test security. Consider a CAT with the following characteristics: (a) item pre-knowledge means a probability of correct response equal to 1; (b) all the examinees start with the same estimated trait level ($\hat{\theta}_0$ is constant); and (c) different examinees with the same estimated trait level receive the same item. Imagine two different scenarios, where source is a previously tested examinee who shares the content of the items he received with a person who will be tested with the same CAT, the recipient:

(1) A high trait level source of information (e. g., $\theta^S = +2$) and low trait level recipient (e. g., $\theta^R = -2$).This condition corresponds to a low overlap rate situation when there is no item disclosure (Way, 1998). The source would give a correct response to the item due to his ability. As source and recipient receive the same first item, the recipient would give a correct response to it, so $\hat{\theta}_1^R = \hat{\theta}_1^S$. As both examinees have the same estimated trait level, they will receive the same item again and both with give correct responses, so $\hat{\theta}_2^R = \hat{\theta}_2^S$. This dynamic will be repeated until the source gives an incorrect response to an item, but the recipient gives a correct response. We will call *h* this item position in the test. Then, $\hat{\theta}_h^R > \hat{\theta}_h^S$ and both $\hat{\theta}_h^R$ and *h* can be very high. As the recipient has an inflated trait score, he will probably miss the next item, so $\hat{\theta}_{h+1}^R < \hat{\theta}_h^R$. The estimated trait level of the recipient will start to decrease until his estimated trait level is equal to some of the (high) estimated trait level of the source, where more pre-known items will be presented. The expected benefit of item pre-knowledge in this condition would be high, although the overlap rate, when there is no bank disclosure, was small.

(2) Both source and recipient of low level (e. g., $\theta^S = \theta^R = -2$). When there is no bank disclosure, this corresponds to a high overlap condition. Both examinees receive the same first item, but the source will miss it, while the recipient will give a correct response. Thus, $\hat{\theta}_1^R > \hat{\theta}_1^S$. Probably, the recipient will give incorrect responses to the next items, slowly reducing his estimated trait level (Rulison & Loken, 2009), until it is equal to some of the (low) estimated trait level of the source. Almost every time that the recipient responds correctly due to pre-knowledge and gets and inflated (although low) ability estimate, he will have no prior information in order to correctly answer the next item. Although the overlap rate when no item disclosure is present was high, a low benefit of item pre-knowledge is expected. The expected benefit of item pre-knowledge does not correspond to the overlap rate when there was no bank disclosure.

In fact, under the three characteristics of the CAT we have described, we could expect the following : (a) all examinees with very high trait levels would receive basically the same items (there is a clear route of items for obtaining a high estimated trait level); and (b) as the first item is the same for all examinees, it is very easy for any examinee who has item pre-knowledge to be exposed to this route. This is the worst situation for a testing agency in the case of bank disclosure.

Taking this into account, we consider that the overlap rate or the distribution of the exposure rates could be of limited value for evaluating CAT security. We propose that other variables could be more useful for evaluating CAT security: (a) the probability of recipient and sources sharing the first items administered; and (b) the overlap rate for high trait levels. Or, in other words, how many routes are available in a CAT in order to reach a high estimated trait level and how easy is it to be exposed to one of these routes from the beginning of the test?

Importantly, the item selection rules with better performance in terms of overlap rate or distribution of exposure rates do not have to be those offering better results in terms of routes leading to high trait estimation. So, commonly used variables for assessing test security in CATs could be misleading. We will now present different item selection rules and show how they differ in these points.

Item selection rules in CATs

Point Fisher information rule

The most commonly applied selection algorithm in CATs is the administration of the item offering maximum Fisher information in the estimated trait level (Lord, 1980):

$$\arg\max_{i \in B_P} I_i(\hat{\theta}), \tag{1}$$

where $I_i(\hat{\theta})$ is the Fisher information of item *i* for the estimated trait level, $\hat{\theta}$ and B_P is the subset of items belonging to the item bank that can be presented to the examinee. If no restriction is active, B_P consists of those items not presented to that examinee. The Fisher information function for the 3 parameter logistic model (Birnbaum, 1968) is calculated according to Equation 2:

$$I_i(\theta) = \frac{2.89 a_i^{\,2}(1-c_i)}{\left(c_i + e^{1.7a_i(\theta-b_i)}\right)\left(1+e^{-1.7a_i(\theta-b_i)}\right)^2}, \tag{2}$$

where a_i is the discrimination parameter, b_i is the locating parameter and c_i is the pseudo-guessing parameter for item *i*.

This rule, which we will call Point Fisher Information (PFI), leads to a high overlap rate and to a highly unbalanced distribution of exposure rates, with some items presented to almost all examinees and many that are never administered.

Alpha-stratified method

The alpha-stratified method (AS - Chang & Ying, 1999) is the alternative rule to PFI that has attracted the most attention in recent years. With this method a much more balanced usage of items is found, leading to an overlap rate close to the minimum

possible, although the cost for this is an increment in the measurement error with respect to PFI.

In the AS method, prior to the administration of a test with a length of *Q* items, the bank is divided into *K* strata. In order to do so, the *n* items in the bank are increasingly ordered in terms of their *a* parameter. The first n/K items in the bank are assigned to the first stratum, the second n/K to the second stratum, and so on. During the administration of the test, the first Q/K items will be selected from the first stratum, the second Q/K items will be selected between the items in the second stratum, and so on. The item selected is the one with the minimum difference in absolute value between $\hat{\theta}$ and its *b* parameter:

$$\arg\min_{i \in B_P} \left| \hat{\theta} - b_i \right|. \tag{3}$$

In this case, B_P consists of the intersection between the non-presented items and the items belonging to the active stratum for that item position in the test.

Progressive method

Another option for reducing the overlap rate, with negligible effects in accuracy in comparison with PFI, is to select items randomly at the beginning of the test and, as the test goes on, to increase the relevance of the Fisher information in the item selection (Barrada, Olea, Ponsoda, & Abad, 2008; Li & Schafer, 2005; Revuelta & Ponsoda, 1998). The progressive method (PG - Revuelta & Ponsoda, 1998) uses this idea. In the PG method, the item selected is the one that maximizes the sum of two elements, one determined by the Fisher information and the other by a random number:

$$\arg\max_{i \in B_P} \left[(1 - W_q) R_i + W_q I_i(\hat{\theta}) \right], \tag{4}$$

where *q* is the item position in the CAT, W_q is the weight of item information in the selection criterion and R_i is a random number belonging to the interval [0, $\max_{i \in B_P} I_i(\hat{\theta})$].

Barrada et al. (2008) have proposed the following equation to relate W_q to *q*:

$$W_q = \begin{cases} 0 & if \quad q = 1 \\ \dfrac{\sum_{f=1}^{q} (f-1)^t}{\sum_{f=1}^{Q} (f-1)^t} & if \quad q \neq 1 \end{cases}. \tag{5}$$

The *t* parameter marks the speed at which the weight of the random component is reduced and, thus, defines the improvement in the overlap rate and accuracy reduction, in comparison with PFI. With *t* equal to 1, it is possible to get an accuracy equivalent to that obtained with the PFI method, while improving the item bank security. Many other item selection rules have been proposed for improving bank security (Georgiadou, Triantafillou, & Economides, 2007). We will restrict our attention to the PFI, AS and PG methods. PFI is, currently, the common standard. AS is an option leading to a major reduction in overlap, with an increment in the measurement error. PG permits a more balanced usage of items, when compared with PFI, while maintaining the accuracy.

Hypothesis

Both the PFI and AS have a negative characteristic for the scenario of bank disclosure: given the same B_P, two examinees with the same estimated trait level will receive the same item. However, when the PG is applied, the same estimated trait level can lead to different items administered, especially in the initial moments of the test, as item selection starts randomly.

From previous studies, it is known that the AS method has an overlap rate lower than the overlap obtained with the PG method and also a more homogeneous distribution of item exposure rates. So, one hypothesis, based on current standards for evaluating test security, would be that the AS method would be better in a condition of bank disclosure. However, we have shown that the PG method outperforms the AS method in the ease of incorporating some of the routes leading to a high estimated trait level. Because of this, our expected results when simulating conditions of bank disclosure are that the PG method will outperform the AS method when there is bank disclosure.

Simulation studies

To check the idea that a lower overlap rate or a more balanced usage of the items do not imply a higher resistance to item disclosure and that the other proposed variables are more adequate, we conducted a series of simulation studies comparing PFI, AS and PG item selection rules. In the first simulation, we present the condition of no item bank disclosure. This study will enable us to distinguish between the common expectation (AS, the method with lower overlap rate and more balanced usage of items, should be the method less affected by bank disclosure) and our expectation that it is more important to consider the number of routes for obtaining a high ability estimate and the ease to incorporate one of this routes (PG outperforms the other two

item selection rules in this respect, so it should be the less affected by bank disclosure). The next three simulation studies are intended to evaluate the resistance of these item selection rules under different conditions of bank disclosure. They are complementary, as they face item bank disclosure from different perspectives. If similar results are obtained from them, this will increase the confidence in our conclusions. In Study 2, we analyze the impact of a different number of sources in overestimation of recipient trait level, evaluating if the common indicators of bank security correctly order the different item selection rules in their resistance to bank disclosure. In Study 3, we study which trait level of sources is more useful for different trait levels of recipients. If overlap rate is a correct indicator of bank security, this trait level should be the one where maximum overlap was found in the condition of no disclosure. In Study 4, we analyze how the measurement error increases the longer the item bank has been in use and we introduce some more realistic conditions of bank disclosure.

Study 1: no item disclosure

Method

Item bank and test length: Ten item banks of 500 items were randomly generated. The distributions for the parameters were: $a \sim N(1.2, 0.25)$; $b \sim N(0, 1)$; $c \sim N(0.25, 0.02)$. Length of the test was set at 25 items.

Item selection rules: We evaluated three different item selection rules: PFI, AS and PG methods. In the AS method, the item bank was stratified in 5 strata and an equal number of items were presented from each strata. For the PG method, the t parameters was fixed at 1 (Equations 5).

Restriction of maximum exposure rate: A common approach to improve bank security is to limit the maximum exposure rate (r^{max}) that no item should surpass. To do this, we used the Sympson-Hetter method (Sympson & Hetter, 1985), with r^{max} equal to 0.25.

Trait level of the simulees: We simulated two different conditions. In the first one, where we obtained results for the overall population, the real trait level of the simulees was randomly extracted from a distribution N(0, 1). For each item selection rule and item bank, 5,000 simulees were sampled. In the second condition, we were interested in the results conditional on trait levels. To do so, we simulated 1,000 examinees for 9 different and equally spaced θ points, ranging from -2 to 2.

Estimation/assignment of trait level: The starting $\hat{\theta}_0$ was randomly selected from the uniform interval (-0.5, 0.5). Maximum-likelihood estimation has no solution in real numbers when there is a constant response pattern, all correct or all incorrect responses. Thus, until there was at least one correct and one incorrect response, $\hat{\theta}$

was assigned using the method proposed by Dodd (1990): when all the responses were correct, $\hat{\theta}$ was increased by ($b_{max} - \hat{\theta}$)/2; if all the responses were incorrect, $\hat{\theta}$ was reduced by ($\hat{\theta} - b_{min}$)/2. After the constant pattern was broken or when the test was finished, we applied maximum-likelihood estimation, with the restriction that $\hat{\theta}$ had to be in the interval [-4, 4].

Performance measures: Five dependent variables were used for the comparison between conditions.

(a) RMSE, calculated according to Equation 6:

$$RMSE = \left(\sum_{g=1}^{v} \left(\hat{\theta}_g - \theta_g \right)^2 / v \right)^{1/2}, \tag{6}$$

where v is the number of simulees;

(b) distribution of item exposure rates ($r_{i,1..Q}$): the exposure rate of item i considering the whole test (from position 1 to position Q, being Q the test length);

(c) overlap rate for the overall population. The common reported value of overlap rate is the pair-wise overlap rate, which provides information about the mean proportion of items shared by two examinees. If an item bank is disclosed, it would be possible for an examinee to gain item pre-knowledge from more than one source. Because of this, we will calculate the overlap rate with z different sources of information. This overlap rate is calculated according to Equation 7:

$$T^{z}_{1..Q,1..Q} = \frac{\sum_{i=1}^{n} r_{i,1..Q} \left[1 - \left(1 - r_{i,1..Q} \right)^{z} \right]}{Q}, \tag{7}$$

where n is the item bank size, z is the number of sources of item information and $T^{z}_{1..Q,1..Q}$ is the overlap rate considering z sources and the whole test. $T^{z}_{1..Q,1..Q}$ provides information about the proportion of items an examinee will face with item pre-knowledge, given z sources. When z is equal to 1, $T^{z}_{1..Q,1..Q}$ is the pair-wise overlap rate. We calculated this overlap rate with z values ranging from 1 to 5 sources.

(d) overlap rate between different trait levels. The mean overlap rate between an examinee with trait level equal to θ_1 and an examinee with trait level equal to θ_2 is

$$T^{z=1,\theta_1,\theta_2}_{1..Q,1..Q} = \frac{\sum_{i=1}^{n} r^{\theta_1}_{i,1..Q} r^{\theta_2}_{i,1..Q}}{Q}, \tag{8}$$

where $r_{i,1..Q}^{\theta_1}$ and $r_{i,1..Q}^{\theta_2}$ refer to the item exposure rate of item *i* for the two sets of trait levels.

(e) probability that an examinee already tested could inform another examinee about the item content for each of the item positions. This probability is equal to the overlap rate, with *z* equal to 1, between a whole test and any single item position ($T_{1..Q,q..q}^{z=1}$):

$$T_{1..Q,q..q}^{z=1} = \sum_{i=1}^{n} r_{i,1..Q} r_{i,q..q} \ , \tag{9}$$

where $r_{i,q..q}$ is the exposure rate of item *i* in just the *q*-th position of the test.

Some studies have reported the overlap rate conditional until an item position in the test ($T_{1..q,1..q}^{z=1}$; Barrada, Velkamp & Olea, 2009). This information, although useful, is not equivalent to the probability of an examinee informing about the item content conditional on item position in the test. Imagine an item bank of 10 items and a test length of 10 items. Applying random item selection, the overlap between examinees until the *q*-th item would be .1**q*, but the probability of an examinee informing another about the any item he will receive is equal to 1.

The first variable measures accuracy. The second and third are the common indicators of test security. The fourth and fifth are the ones that we hypothesize can measure better test security.

Results

The results for the overall population in terms of overlap rate and RMSE can be seen in Table 1. As expected, the selection rule with the greatest overlap rates was PFI. With the AS method, much lower overlap rates were obtained. The PG method was between these extremes. The order in overlap rates is the same for the three methods for any number of sources considered. While the PFI and PG methods offered the same RMSE, the accuracy was reduced with the AS method, as we obtained an RMSE 0.07 higher than with the other two methods.

Figure 1 shows the exposure rates of the items considering the whole population. In accordance with the overlap data, the exposure rates for the PFI method are the most unbalanced, while those for the AS method are the most homogeneous. The distribution of the PG method was located between the PFI and AS methods. With the PFI method, up to 57% of the items had an exposure rate equal to 0. With the AS method, the proportion of unused items is 1%. For the PG method, there were no items that were never presented. The proportion of items with exposure rates over or near r^{max} (> 0.2) were 12% for the PFI method, 1% for the AS method, and 7% for the PG

method. With the Sympson-Hetter method, some items had exposure rates slightly over r^{max} (van der Linden, 2003).

Table 1

Overlap rate and RMSE according to item selection rule with no bank disclosure.

	$T^1_{1..Q,1..Q}$	$T^2_{1..Q,1..Q}$	$T^3_{1..Q,1..Q}$	$T^4_{1..Q,1..Q}$	$T^5_{1..Q,1..Q}$	RMSE
PFI	.19	.35	.46	.56	.63	.26
AS	.07	.14	.20	.26	.31	.33
PG	.14	.25	.34	.41	.47	.26

Table 2

Overlap rate according to trait levels and item selection rules. Bold faced figures correspond to trait level where the overlap rate is greater given a fixed overlap rate (maximum by row).

			trait level								
			-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
PFI	trait level	-2	**.61**								
		-1.5	.47	**.54**							
		-1	.22	.38	**.50**						
		-0.5	.10	.16	.32	**.43**					
		0	.06	.07	.13	.27	**.40**				
		0.5	.04	.05	.06	.11	.27	**.43**			
		1	.03	.03	.04	.06	.13	.31	**.50**		
		1.5	.03	.03	.03	.05	.08	.15	.37	**.59**	
		2	.02	.02	.03	.04	.06	.09	.20	.49	**.69**
AS	trait level	-2	**.44**								
		-1.5	.29	**.27**							
		-1	.12	.15	**.16**						
		-0.5	.05	.06	.10	**.11**					
		0	.03	.03	.05	.08	**.11**				
		0.5	.02	.02	.03	.05	.08	**.13**			
		1	.02	.02	.02	.03	.06	.11	**.19**		
		1.5	.02	.02	.02	.03	.05	.08	.18	**.31**	
		2	.01	.02	.02	.03	.04	.07	.15	.35	**.59**
PG	trait level	-2	**.44**								
		-1.5	.34	**.38**							
		-1	.17	.28	**.35**						
		-0.5	.07	.12	.23	**.31**					
		0	.03	.04	.09	.20	**.29**				
		0.5	.02	.02	.04	.08	.20	**.31**			
		1	.02	.02	.02	.03	.08	.23	**.36**		
		1.5	.02	.02	.02	.02	.04	.11	.27	**.41**	
		2	.02	.02	.02	.02	.03	.05	.14	.34	**.48**

In Table 2, the overlap rates between examinees of the same and different trait levels are shown. The higher overlap rates are between examinees of the same trait level. The higher overlap for examinees of the same trait level can be found for examinees with extreme traits, as fewer items are available there. The higher the difference between the trait levels of the examinees, the lower the overlap (Way, 1998). The

higher overlap rates correspond to the PFI method. The PG method reduces the overlap. In general, the AS is the method with the lowest overlap. An exception can be

Figure 1
Exposure rates of items according to item selection rule with no bank disclosure. Items ordered according to their exposure rates.

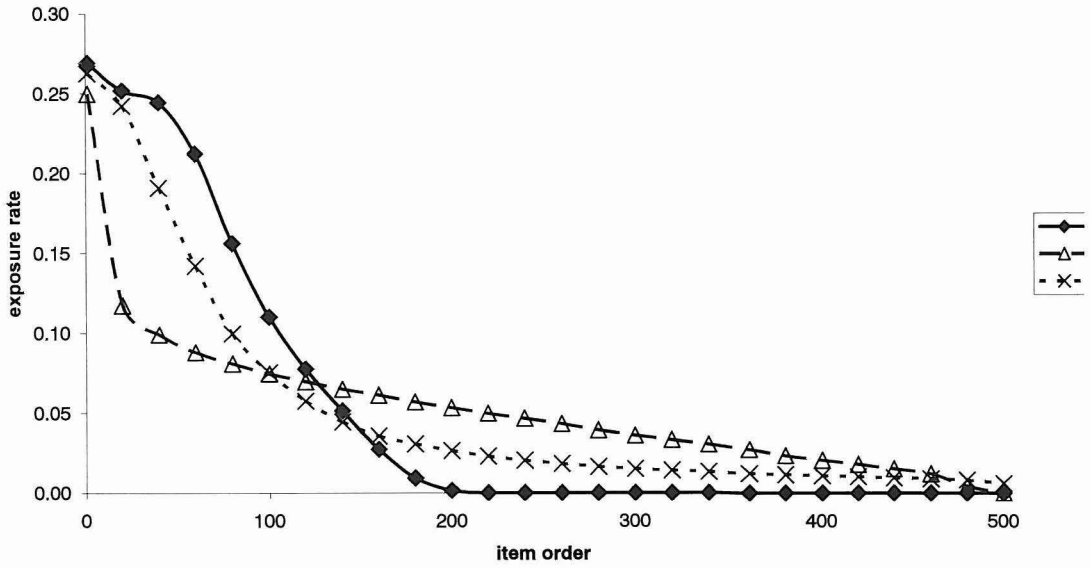
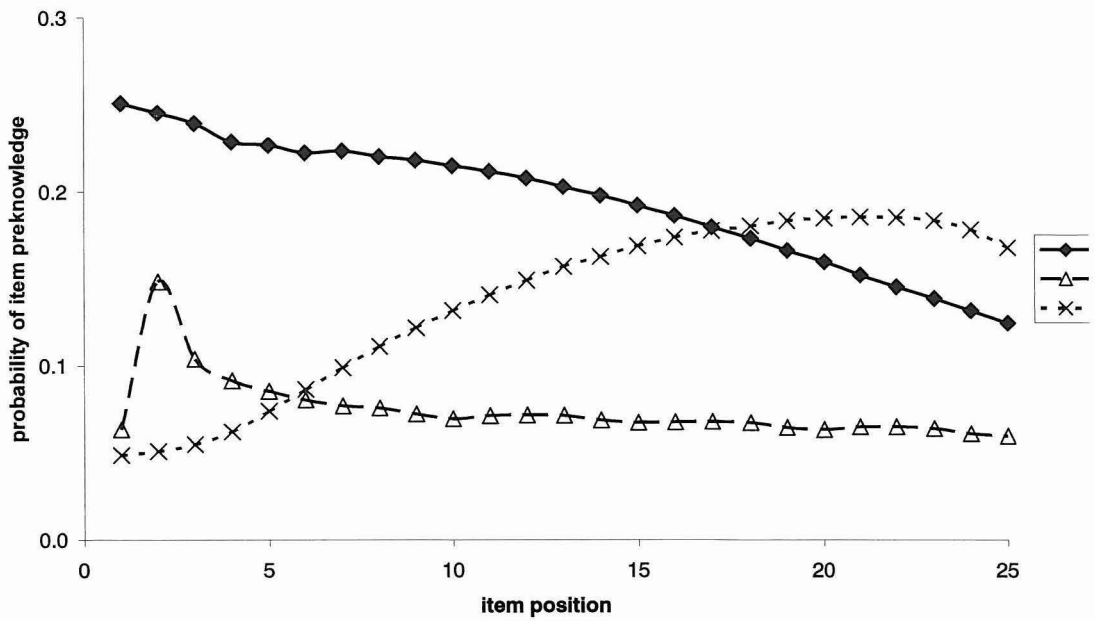


Figure 2
Probability of item preknowledge according to item position and item selection rule with no bank disclosure.



seen in the overlap between examinees with trait levels equal to 2, where it is higher for the AS method than for the PG method.

In Figure 2 shows the probability of an examinee providing information about the item content at each item position. For the PFI method, the probability of an examinee giving information about the content of the items reduces as the test goes on. The most interesting result is the pattern of results that is found when comparing the AS and PG methods. For the PG method, because of the randomness in item selection at the beginning of the test, the probability of receiving information about item content is lowest at the start of the test and increases with each new item. For the AS method, the probability of receiving information is higher at the start of the test and decreases as the test continues. After the sixth item, the probability of receiving information is lower for the AS method than for the PG method. Another interesting result is that, for the PG method, the probability of receiving information about the content of an item at the end of the test is greater than for PFI. This is because with PG, the most likely to be selected items are not presented until the end of the test.

It has to be noted that the probabilities of Figure 2 hold while no examinee has item pre-knowledge or until the examinee reaches the item position where there is item pre-knowledge. With item pre-knowledge, as the probability of a correct response changes with respect to the simulation, these probabilities would be changed.

Discussion

In this study, we have performed simulations to evaluate three different item selection rules in terms of their item exposure control. According to Tables 1 and 2 and Figure 1, the selection rule which seems to be preferred in order to maximize test security is the AS method, as it has lower overlap rates for the overall population of examinees, a more balanced distribution of the exposure rates and, in general, lower overlap rates conditional on trait levels.

The most generally accepted assumption would be that the AS method would be the method with the lowest inflation of the trait levels when there was item bank disclosure. However, our expectation is that the PG method will outperform the AS method in this condition. We have argued above that especial attention should be paid to two other variables: (a) the number of different paths available for obtaining an estimate of a high trait level; and (b) the ease of incorporating one of these paths from the beginning of the test. Table 2 and Figure 2 show that the risks for these two points are higher for the AS method than for the PG method. For the AS method, when compared with the PG method, the overlap rate conditional on trait level equal to 2 is greater and the

probability of being informed about the item content at the beginning of the test is higher.

To check the hypothesis that the common indicators of test security could be misleading and that a CAT using PG would be more robust to item disclosure than a CAT using AS, we conducted the following three studies.

Study 2: effect of disclosure according to the number of sources

Method

The method of this second study is similar to that of the first, except in the points which we describe now. In this study, we investigated the effect on overestimation of the trait level when examinees-recipients take the exam after contacting m independent examinees-sources of items. The number of sources could be 1, 2... up to 10. The process was: (a) to simulate the exam for m sources as standard CATs; (b) to do the union of the different sets of items and fix the probability of correct response to these items as equal to 1 for the recipient; and (c) to simulate the CAT for the recipient with these changed probabilities for some items. Sources and recipients were extracted from a standard normal distribution. For each of the 10 conditions (number of sources) 5,000 replicas were simulated. Four dependent variables were recorded: (a) proportion of pre-known items in the bank (cardinal of the set formed by the union of items presented to the different sources divided by item bank size); (b) proportion of pre-known items in the test (cardinal of the set formed by intersection of the pre-known items in the bank and the items presented to a recipient divided by test length); (c) bias; and (d) RMSE. The bias was calculated according to Equation 10:

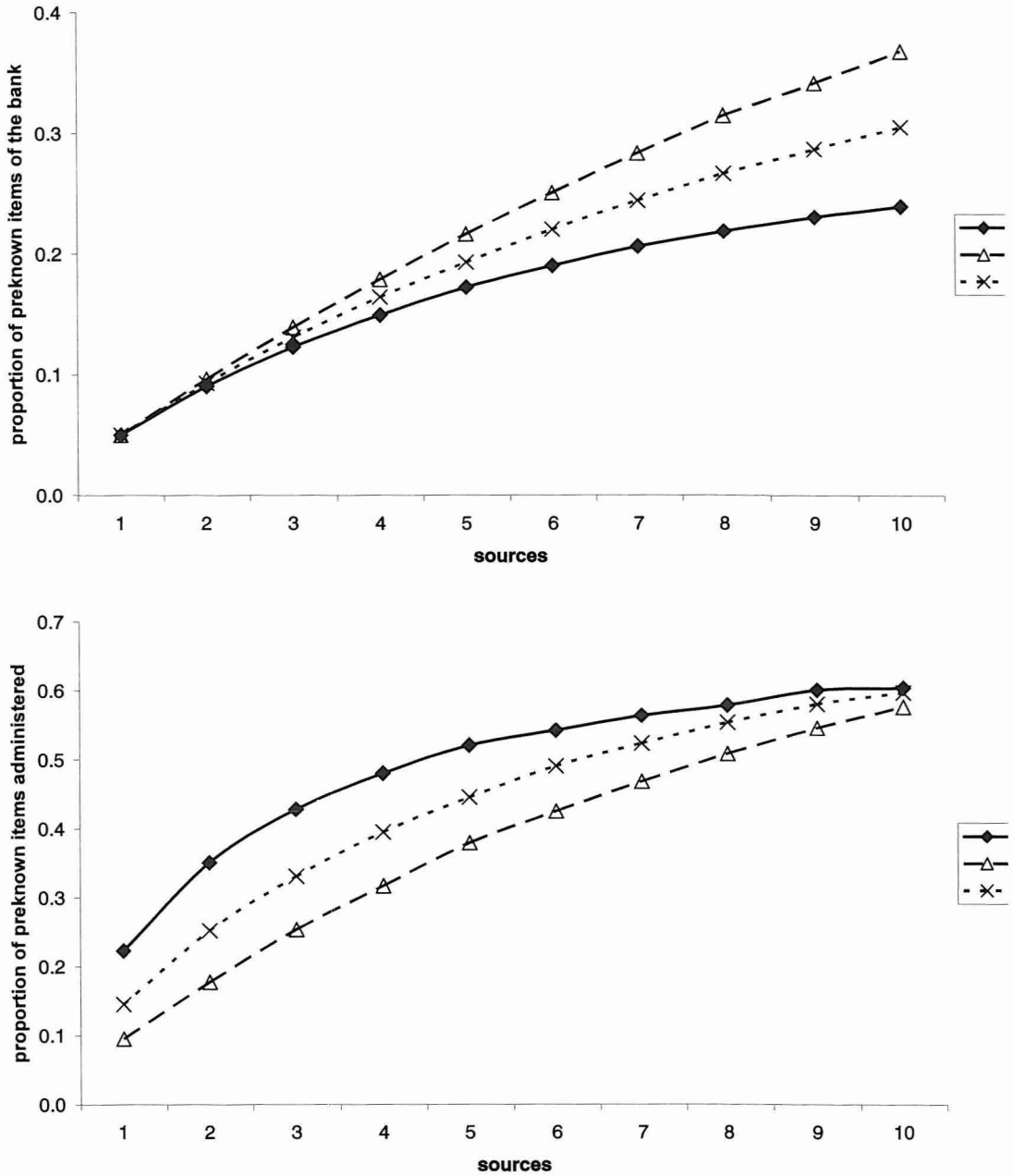
$$Bias = \sum_{g=1}^{v} \left(\hat{\theta}_g - \theta_g\right) / v \tag{10}$$

Results

The upper panel of Figure 3 shows the proportion of pre-known items in the bank according to the number of sources. The higher the number of sources, the higher this proportion is. With a high overlap rate, as evaluated in Study 1, the sources offer redundant information. Because of this, the higher the overlap rate, the lower the proportion of pre-known items in the bank. Thus, the item selection rule leading to higher knowledge of the bank is the AS method.

The proportion of pre-known items administered in the test can be seen in the lower panel of Figure 3. Again, the higher the number of sources, the higher this proportion is. The higher the overlap rate, the higher the proportion of items in the test for which there is pre-knowledge. The method for which the highest proportion of pre-known

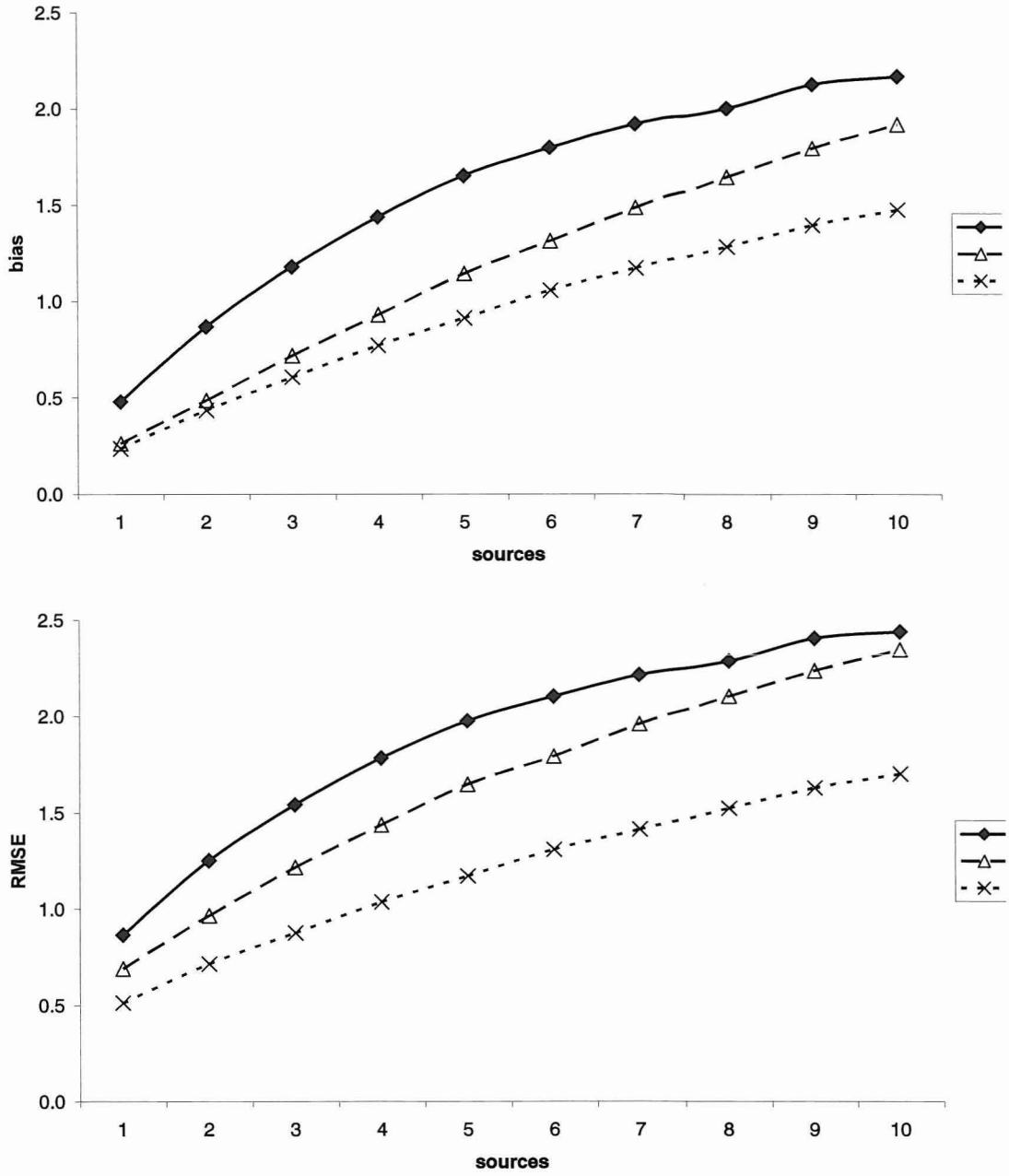
Figure 3
Proportion of preknown items in the item bank and proportion of preknown items administered according to the number of sources and item selection rule.



items is administered is the PFI method; the one with the lower proportion is the AS method. The results of this figure and the previous one follow what could be expected from the results of Study 1.

Figure 4 shows the bias (overestimation) and RMSE when item disclosure is present. The selection with worse resistance to item disclosure is PFI. Contrary to the

Figure 4
Bias and RMSE according to the number of sources and item selection rule.



hypotheses derived from the overlap rate and item exposure rates in Study 1, the next is the AS method. The method that is less affected by bank disclosure is the PG method. The presence of each new source produces an increment in bias and RMSE that is higher for the AS method than for the PG method. In any case, the presence of just one source leads to bias and RMSE to values that could be considered unacceptable.

Discussion

Two results from this study deserve special attention. First, the method with the lower overlap rate, the AS method, is not the one with the lower impact of item disclosure. Second, the method with the lower proportion of pre-known items in the test, the AS method again, is not the method with the greater resistance to bank disclosure. As we hypothesized above, the PG method is the one offering greater security when the item bank is disclosed.

The AS method has two problems in terms of resisting item pre-knowledge: (a) there are not many different paths for the estimation of high trait levels; and (b) it is easier than with the PG method to incorporate one of these paths from the beginning of the test. In this way, we can explain why with the AS method we find a higher bias and RMSE when item disclosure is simulated.

It should be noted that we have not chosen the worse conditions when simulating the AS method. Randomly assigning the initial trait level in the interval (-0.5, 0.5), as we have done, reduces the probability of item pre-knowledge of the first administered items for the AS and PFI methods, while it does not affect this probability for the PG method, as it starts with random selection. The common practice of starting the CAT with a fixed trait level would deteriorate the performance of the PFI and AS methods, while not affecting the PG method.

That higher impact is attained with fewer items with pre-knowledge is probably due to the fact that the items with pre-knowledge are mainly situated at the beginning of the test. When this happens, trait estimation is highly shifted to the positive extreme, and many items are required to reduce the overestimated trait level and many items are administered providing low information at the real trait level. When the pre-known items are presented at later stages of the test, as with the PG method, the likelihood-function is more peaked, so the overestimation due to pre-known items is smaller.

In Study 2, we have evaluated the effects of item disclosure from random sources. Another approach, taken in Study 3, is to evaluate the impact of different trait levels of the sources and recipients.

Study 3: effect of disclosure according to the trait level of the sources and the recipients

Method

The method of this study is equivalent to that employed in the previous studies, except in the following points. The source and recipient trait levels were manipulated with 9 levels, from -2 to 2 with increments of 0.5. Each recipient had just one source. Each

Table 3

Proportion of preknown items according to trait levels of the source and the recipient and the item selection rule. Bold faced figures correspond to the trait levels of the source where the proportion of preknown items is greater with the recipient (maximum by row).

			source trait level								
			-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
PFI	recipient trait level	-2	.35	**.38**	**.38**	.28	.22	.17	.14	.13	.13
		-1.5	.26	.33	**.39**	.32	.23	.18	.16	.14	.13
		-1	.15	.24	.33	**.37**	.29	.20	.17	.16	.16
		-0.5	.08	.13	.22	.31	**.35**	.26	.21	.21	.20
		0	.06	.07	.09	.19	.30	**.35**	.32	.31	.29
		0.5	.05	.05	.05	.09	.19	.33	.40	**.41**	.39
		1	.03	.03	.04	.07	.11	.22	.37	.48	**.51**
		1.5	.03	.03	.03	.04	.06	.12	.27	.46	**.58**
		2	.02	.02	.03	.04	.05	.08	.16	.34	**.55**
AS	recipient trait level	-2	**.27**	.23	.16	.07	.05	.04	.05	.07	.11
		-1.5	.17	**.19**	.17	.09	.05	.05	.05	.08	.11
		-1	.08	.11	.13	.11	.07	.06	.06	.10	**.14**
		-0.5	.04	.05	.08	.10	.10	.08	.09	.12	**.17**
		0	.02	.03	.05	.07	.11	.11	.12	.17	**.23**
		0.5	.02	.02	.03	.04	.07	.13	.17	.23	**.34**
		1	.02	.02	.02	.03	.06	.11	.20	.32	**.47**
		1.5	.01	.02	.02	.03	.04	.07	.15	.33	**.60**
		2	.02	.02	.02	.03	.04	.06	.14	.36	**.63**
PG	recipient trait level	-2	.31	**.33**	.27	.13	.06	.04	.03	.03	.02
		-1.5	.23	.29	**.31**	.20	.09	.04	.03	.03	.02
		-1	.13	.22	**.29**	.27	.15	.07	.04	.03	.03
		-0.5	.06	.09	.18	**.25**	**.25**	.16	.07	.05	.03
		0	.03	.04	.08	.16	**.26**	.25	.16	.10	.07
		0.5	.02	.02	.04	.07	.15	.27	**.29**	.21	.14
		1	.02	.02	.02	.03	.08	.18	.30	**.34**	.30
		1.5	.02	.02	.02	.02	.03	.09	.20	.36	**.42**
		2	.02	.02	.02	.02	.03	.06	.12	.28	**.42**

one of the 81 conditions (9 level of source * 9 levels of recipient) was simulated 1,000 times. As dependent variables we used: (a) the proportion of items of the test which were pre-known; and (b) the bias.

Results

In Table 3, the proportion of pre-known items according to trait levels of sources and recipients is shown. This table should be compared with Table 2, where the overlap rate between trait levels in the condition of no bank disclosure was shown. While Table 2 was symmetric, Table 3 is not: a high-level source giving information to a low-level recipient does not have the same effect as the opposite case. In Table 2, the maximum overlap rate was on the diagonal, that is, between examinees of the same trait level. In Table 3, the pattern of results depends on item selection rule. For the PFI method, the maximum proportion of pre-known items is achieved with a source with a trait level 0.5

Table 4

Bias according to trait levels of the source and the recipient and the item selection rule. Bold faced figures correspond to the trait levels of the source that produce the greater gain in the estimated trait levels of the receptor (maximum by row). In grey, the cells where the bias is equal or over .5.

			source trait level								
			-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
PFI	recipient trait level	-2	.76	.96	**1.15**	1.08	1.05	.99	.89	.89	.86
		-1.5	.39	.58	.84	**.89**	.82	.83	.88	.84	.80
		-1	.17	.31	.48	.72	.73	.71	.76	**.83**	.81
		-0.5	.10	.13	.22	.43	.65	.67	.75	**.88**	.86
		0	.09	.09	.11	.23	.43	.67	.84	1.03	**1.06**
		0.5	.07	.07	.06	.10	.20	.47	.77	1.02	**1.19**
		1	.04	.04	.06	.07	.12	.24	.52	.91	**1.23**
		1.5	.04	.03	.03	.05	.07	.10	.28	.63	**1.07**
		2	.02	.04	.04	.04	.05	.09	.13	.33	**.79**
AS	recipient trait level	-2	.61	.61	.50	.30	.28	.31	.41	.49	**.74**
		-1.5	.30	.36	.42	.32	.26	.31	.38	.58	**.66**
		-1	.13	.18	.28	.29	.24	.30	.35	.55	**.74**
		-0.5	.10	.11	.14	.20	.27	.32	.41	.55	**.77**
		0	.06	.07	.12	.13	.22	.30	.41	.64	**.92**
		0.5	.06	.07	.09	.09	.14	.27	.41	.67	**1.14**
		1	.04	.06	.06	.06	.11	.16	.37	.72	**1.26**
		1.5	.03	.03	.05	.04	.04	.09	.20	.54	**1.30**
		2	.03	.03	.03	.05	.04	.07	.14	.46	**1.05**
PG	recipient trait level	-2	.63	**.78**	.74	.40	.22	.17	.12	.10	.07
		-1.5	.31	.47	**.63**	.48	.28	.14	.12	.11	.06
		-1	.12	.23	.38	**.49**	.34	.20	.13	.11	.11
		-0.5	.04	.08	.17	.31	**.41**	.35	.21	.16	.12
		0	.04	.05	.06	.15	.32	**.43**	.37	.30	.21
		0.5	.04	.03	.05	.06	.13	.34	**.51**	.50	.41
		1	.03	.03	.04	.03	.07	.17	.41	.65	**.70**
		1.5	.02	.02	.03	.04	.04	.08	.20	.52	**.81**
		2	.05	.03	.03	.03	.04	.06	.10	.31	**.67**

or 1 points above the recipient. For the AS method, except for low trait level recipients, the maximum pre-known items are offered by high-level sources. For the PG method, the maximum proportion is given by sources with equal or slightly greater (+0.5) trait levels than the recipients.

In Table 4 we present the bias (overestimation) according to trait levels of sources and recipients. The pattern of results is markedly different between the PFI and AS methods, on the one hand, and the PG method on the other. For both the PFI and AS methods (with some exceptions in the PFI for low level recipients), the higher the trait level of the sources, the higher the overestimation. For the PG method, the higher overestimation comes from gaining information from a source slightly above (+0.5 or +1) the trait level of the recipient.

Discussion

If you are an examinee seeking to inflate your score by means of item pre-knowledge, which trait level source should you look for? According to studies where no item disclosure was simulated, that examinee should try to find an examinee with a similar trait level. When item disclosure is simulated, the answer changes and depends on the item selection rule implemented in the CAT. Both PFI and AS offer the same answer (with slight exceptions): in general, look for a source with a high trait level. Or, in other words, one source fits all the recipients. The source trait level leading to higher benefit is not the one with the higher overlap rate when there is no bank disclosure. If the CAT uses the PG method, examinees trying to boost their estimated trait level should try to find sources with slightly higher trait levels. Again, overlap rate, as shown in Study 1, would lead to incorrect predictions.

In Studies 2 and 3,(a) all the examinees had item pre-knowledge; (b) the sources could give perfect information about the items they received; (c) the recipients could remember perfectly all the items the sources shared with them; and (d) item pre-knowledge was equal to a probability of a correct response equal to 1. Although all of these conditions were useful for capturing how item disclosure works, in the next study we present a more realistic simulation. Our goal was to check whether, when changing the four points noted above, the pattern of results still holds.

Study 4: effect of disclosure according to examinee position in the item bank life

Method

In this study, the longer the item bank has been in use, the higher the probability of an examinee knowing one or several sources. We set the probability of an examinee knowing each previously examined person as equal to 0.001. For the *(h+1)*-th examinee, for each of the *h* previous examinees, a random number was extracted from a uniform distribution (0, 1). Only if the number was lower than 0.001, did that examinee become a source. The probability of the source giving information about each single item he received was equal to 0.15. Whenever he shared the item, the probability of a correct response to that item was fixed at 1 for the recipient. The probability of the source sharing the content of each item can also be viewed, in this context, as the probability of the recipient remembering it. Clearly, the probabilities chosen are arbitrary, but serve to show the effect of bank disclosure under different conditions from those simulated previously. Unreported simulations with different values lead to equivalent patterns of results. As dependent variables, we will present

the number of pre-known items in the test, the bias, and the RMSE. To improve the clarity of the figures, we show the results averaged for each 200 examinees.

Results

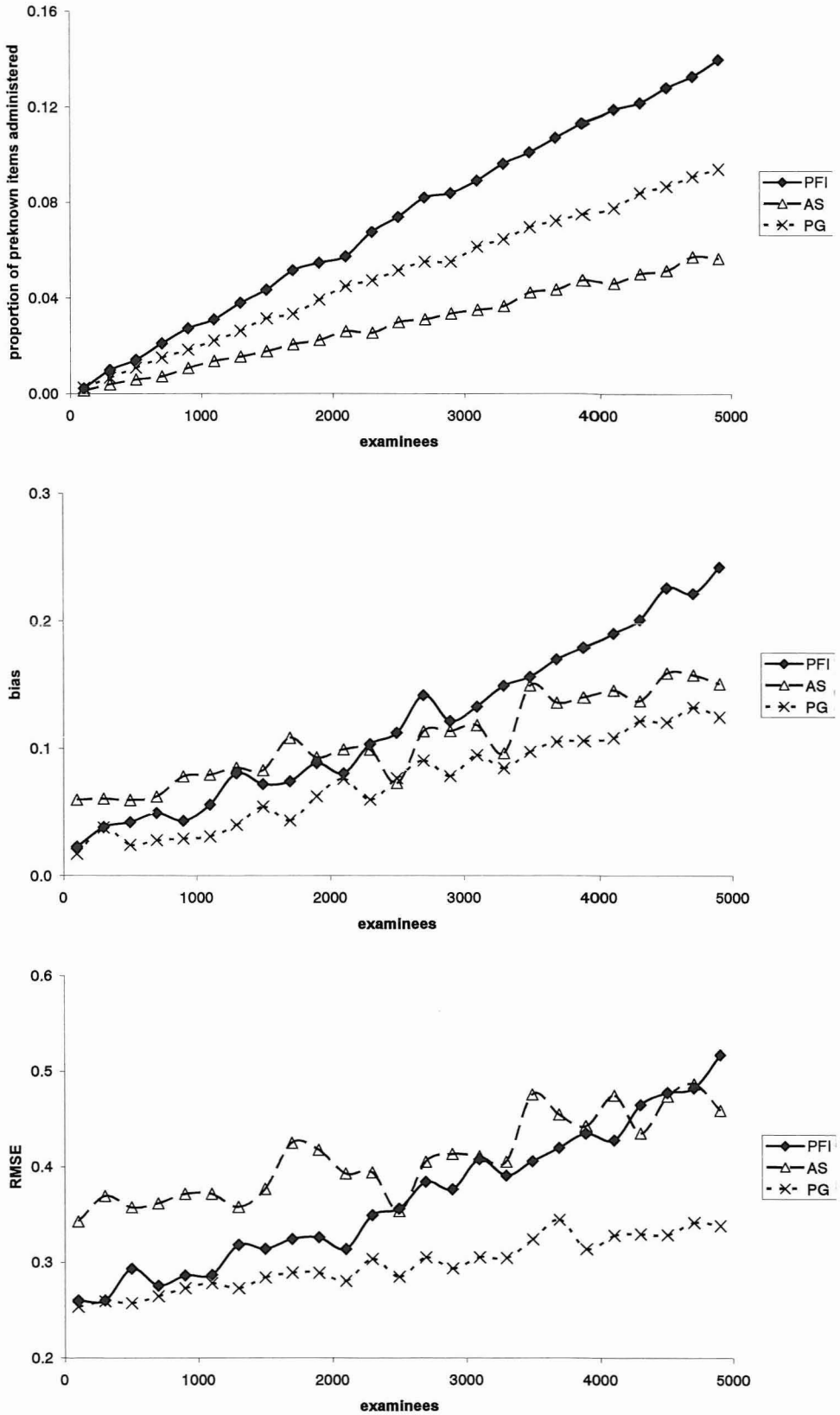
As expected due to the simulation procedure, the higher the examinee position in the life time of the item bank, the higher the number of pre-known items in the test. This can be seen in the upper panel of Figure 5. Consistent with the results reported in Figure 3, the item selection rule which leads to the higher number of items administered with prior information is the PFI method. The method where fewer pre-known items are administered is the AS method. The PG method is between these two extremes.

Middle and lower panels of Figure 5 show how these pre-known items affect the accuracy of trait level estimation. The later the position of the examinee, the higher both the bias and the RMSE, as the expected number of sources increases. Consistent with Table 1, at the beginning of the item bank life the PFI method has a lower measurement error than the AS method and a bias and RMSE equivalent to the ones obtained with the PG method. But, as the PFI method is the item selection rule where the measurement error increases more rapidly, when 2,400 examinees or more have been tested, its bias is higher than with the AS method and for the last thousand of examinees its RMSE is also higher. For the AS and PG methods, the speed with which they increment the measurement error appears equivalent. As the PG method started with higher accuracy, this method shows throughout the bank life lower bias and RMSE than the AS method.

Discussion

In this study, we have changed the way in which item disclosure is simulated. In this case, the sources can also have an inflated trait level. Neither the source nor the recipient perfectly memorizes the content of the item. In these different conditions, again, the item selection rule with lower resistance to bank disclosure is the PFI method. The order of the different item selection rules in terms of accuracy depends on the length of the item bank life. At the beginning, the worst method is AS; after some examinees, it is the PFI method. Importantly, for any examinee position, the method with the lower measurement error is the PG method. As in Study 2, we found that administering a lower number of pre-known items, as the AS method does, does not lead to a lower effect of overestimation in the trait levels.

Figure 5
Proportion of preknown items administered, bias and RMSE according to examinee position and item selection rule.



Conclusions

The purpose of these four studies was to evaluate the validity of two variables that are frequently used when checking test security in CATs: distribution of the item exposure rates and overlap rate. It has been commonly assumed that the more balanced the distribution and the lower the overlap rate, the higher would be the resistance to item disclosure. We have argued that this idea could be incorrect. First, we have shown the results of what could be considered a typical study (Study 1): in the condition of no disclosure, the conclusion should be that the AS method is the safer one. In the following three studies, we have shown that this conclusion does not hold. The AS method, when compared with the PG method: (a) is more affected by the presence of sources and any new source increases the bias and RMSE more (Study 2); (b) has a 'golden source', the one with a high trait level that will inflate the trait estimation of all the recipients the most (Study 3) and this 'golden source' is not the one with higher overlap as measured in Study 1; and (c) is never, when considering the item bank life under conditions of bank disclosure, the item selection rule to be preferred in terms of accuracy (Study 4). Taking all these facts into account, we can conclude that the usual variables reported in the extensive literature on test security and item exposure control in CATs should be considered with caution.

We find another interesting result: a higher number of pre-known items does not lead directly to a higher effect on the accuracy of estimation. The effect of bank disclosure cannot be tested by the overlap rate, the distribution of item exposure rates, the percentage of the item bank that is pre-known, or by the percentage of the items administered during the test that are pre-known. It seems that the only way to detect the safer item selection rules is by carrying out simulations where item disclosure is simulated.

What seems clear is that one way of improving test security is to increase randomness in item selection at the beginning of the test, as the PG method does. Thus, we reduce the overlap rate when the test starts and increase the number of possible combinations of items leading to an estimation of high trait levels. Several other options that could improve the resistance to bank disclosure could also be considered. For instance, the testing agency could construct an item bank with a higher mean or standard deviation of the *b* parameter distribution, so more items could be available at the high extreme and, thus, probably reducing the overlap rate at the high levels. Another option would be to use methods for the restriction of a maximum rate conditional on trait levels

(Stocking & Lewis, 2000; van der Linden & Veldkamp, 2007), reducing the r^{max} value especially for high trait levels.

In these studies, we have not used any method to try to detect the examinees with item pre-knowledge by means of their pattern of responses (Bradlow, Weiss & Cho, 1998; McLeod & Lewis, 1999; McLeod et al., 2003; Nering, 1997; van Krimpen-Stoop & Meijer, 2001) or by means of their response times (van der Linden & van Krimpen-Stoop, 2003; van der Linden & Guo, 2008). Clearly, these lines of research are useful, but, perhaps, problematic in practice. What should a testing agency do with an examinee that probably has item pre-knowledge? As the evidence is only probabilistic, it is hard to believe that any examinee could fail the exam or be obliged to repeat the exam with this evidence. Segall (2004) presents an interesting idea: instead of making a decision at the end of the test, it could be possible to adapt the items to be presented, switching to infrequently exposed items when pre-knowledge is suspected.

Our approach is different. Instead of detecting recipients, we try to identify the item selection rule that would produce the lowest benefit for recipients. The lower the benefit, the lower the probability of an examinee spending time trying to find a source.

The algorithm we have employed adapts after each item administered the next item to be presented. Another option is to reduce the number of times of selection and selecting predefined packages of items, as done with multistage testing (Luecht & Nungester, 1998). With this option and considering how item packages are currently built (e. g., Belov & Armstrong, 2008; Breithaupt & Hare, 2007), the probability of item pre-knowledge at beginning of the test is necessarily greater than the probability with the PG method. It remains for future research to study the differential effect of item disclosure for CAT and multistage testing.

References

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness to the Fisher information for improving item exposure control in CATs. *British Journal of Mathematical and Statistical Psychology, 61*, 493-513.

Barrada, J. R., Veldkamp, B. P., & Olea, J. (2009). Multiple maximum exposure rates in computerized adaptive testing. *Applied Psychological Measurement, 33*, 58-73.

Belov, D.I., & Armstrong, R. D. (2008). A Monte Carlo approach to the design, assembly, and evaluation of multistage adaptive tests. *Applied Psychological Measurement, 32*, 119-137.

Birnbaum, A. (1968). Some latent ability models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.) *Statistical theories of mental test scores* (pp. 392-479). Reading, MA: Addison-Wesley.

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93*, 910–919.

Breithaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement, 67*, 5-20.

Chang, H. H.. (2004). Understanding computerized adaptive testing – From Robbins-Monro to Lord and beyond. In David Kaplan (Ed.) *The SAGE handbook of quantitative methodology for the social sciences* (pp. 117-133). Thousand Oaks, CA: Sage Publications.

Chang, H. H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

Chang, H. H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67*, 387-398.

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*, 129-145.

Davey, T., & Nering, N. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward, (Eds). *Computer-based testing: Building the foundation for future assesments* (pp. 165-191). Mahwah, NJ: Lawrence Erlbaum.

Dodd, B. G. (1990) The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14*, 355-366.

Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8). Retrieved February 17, 2009, from http://escholarship.bc.edu/jtla/vol5/8/.

Li, Y. H., & Schafer, W. D. (2005). Increasing the homogeneity of CAT's item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. *Journal of Educational Measurement, 42*, 245-269.

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems*. New Jersey: Lawrence Erlbaum.

Luecht, R. M., & Nungester R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 229-249.

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23*, 147-160.

McLeod, L. D., Lewis, C, & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27*, 121-137.

Mills, G. N., & Steffen, M. (2000). The GRE computer adaptive test: Operation issues. In W. J. van der Linden and C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75-100). Boston: Kluwer Academic Press.

Nering M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 115-127.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Rulison, K. L., & Loken. E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33*, 83-101.

Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics, 27*, 163–179.

Segall, D. O. (2004). A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 29*, 439-460.

Stocking, M. L., & Lewis, C. L. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice* (pp. 163-182). Dordrecht, The Netherlands: Kluwer Academic.

Stocking, M. L., Ward, W. C., & Potenza, M. T. (1998). Simulating the use of disclosed items in computerized adaptive testing. *Journal of Educational Measurement, 35*, 48-68.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military*

Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W. J. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 28*, 249-265.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*, 365-384.

van der Linden, W.J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika, 68*, 251-265.

van der Linden, W. J., & Veldkamp, B. P. (2007).Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics 32,* 398-418.

van Krimpen-Stoop, E. M. L. A., & Meijer R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*, 199-217.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17-27.

Yi, Q., Zhang, J., & Chang, H. H. (2008). Severity of organized item theft in computerized adaptive testing: A simulation study. *Applied Psychological Measurement, 32,* 543-558.

PARTE III – DISCUSIÓN Y CONCLUSIONES

Capítulo 7.
Discusión y conclusiones

7.1. Discusión

En el Capítulo 1 describíamos los pasos para la selección de items en un TAI. Señalábamos que el primer paso es determinar un subconjunto del banco de items, B_q, del que en un paso posterior se elige un ítem. Los dos primeros estudios han buscado ofrecer mejoras para cada uno de estos dos pasos del proceso de selección. Los estudios tercero y cuarto han querido mostrar métodos mejores y más robustos para la comparación de reglas de selección. A continuación, presentamos una discusión general de los diferentes estudios.

El método de Múltiples Tasas Máximas (MRM, por *multiple* r^{max}), basado en el método de elegibilidad del ítem (van der Linden & Veldkamp, 2004), comparte con éste la capacidad para adaptar los parámetros de control de la exposición tras cada examinado. De este modo, y a diferencia del método Sympson-Hetter (Sympson & Hetter, 1985), el más empleado hasta el momento, no se requieren simulaciones previas, se controla mejor la tasa máxima de exposición y el funcionamiento del método no es vulnerable a discrepancias entre la distribución de niveles de rasgo simulada y la real.

Ahora bien, el método MRM representa, en comparación con las alternativas previas para el control de la tasa máxima de exposición, un cambio de modelo. Hay razones tanto teóricas como fundamentadas en estudios previos (Li & Schafer, 2005; Revuelta & Ponsoda, 1998) para defender que presentar items con un bajo nivel de discriminación al comienzo del test no deteriora la recuperación del nivel de rasgo, mientras que de este modo se mejora la seguridad. Un modo para conseguir este resultado es planteando tantas tasas máximas de exposición como items a administrar. El test comienza con un control próximo al máximo posible y este control se va relajando según avanza el test, hasta aproximarse al nivel de control obtenido con los métodos clásicos.

En el estudio presentado hemos mostrado las ecuaciones que permiten implementar este método en un TAI. Igualmente, hemos ofrecido una función que permite controlar mediante un parámetro la velocidad con la que se cambia desde control máximo a

control equivalente al aportado por los métodos previos. Al poner a prueba la propuesta, encontramos que el método es viable y consigue el objetivo de variar la tasa máxima de exposición de un modo creciente a lo largo del test. Los resultados indican que las diferencias en precisión son mínimas o nulas al emplear el método MRM y compararlo con el método de elegibilidad del ítem, mientras que se mejora importantemente la seguridad del banco.

El estudio sobre los métodos progresivo y proporcional comparte una parte importante de la lógica y de los resultados expuestos para el método MRM. En este caso, en lugar de buscar mejorar la seguridad del banco mediante restricciones a B_q, proponemos para este fin cambios en la función de valoración de items. Por un lado, extendemos la investigación realizada con el método progresivo. Hasta el momento, con este método se había estudiado: (a) una transición lineal desde selección aleatoria hasta selección basada por completo en la información de Fisher (Revuelta & Ponsoda, 1998); o (b) una reducción del peso del azar en la selección, haciendo que el método se aplicara únicamente a una parte inicial y menor del test (Eggen, 2001). Puesto que los resultados originales mostraban un deterioro despreciable en precisión cuando se aplicaba el método progresivo, parecía razonable intentar ampliar el peso del azar en la selección, en lugar de limitarlo. Para ello, presentamos una fórmula que, al igual que ocurre con una similar ofrecida para el método MRM, permite controlar a través de un parámetro la importancia concedida a la aleatoriedad en la selección para los distintos items del test.

Por otro lado, desarrollamos el método proporcional, una propuesta que apenas ha recibido atención en la investigación en el campo. Los métodos de selección de items habituales están basados en la optimización de un criterio. Al actuar así, aquellos items que no figuran entre los Q items óptimos para ningún nivel de rasgo (siendo Q la longitud del test) jamás son presentados. Todos los items por debajo de los Q óptimos son tratados como iguales, con independencia de sus diferencias en propiedades psicométricas, y, en términos prácticos, podrían ser eliminados del banco de items. El método proporcional utiliza las funciones de valoración de items para determinar la probabilidad de selección. Al actuar así, se permite que aquellos items que más aportan para la estimación del nivel de rasgo tengan más opciones de ser administrados, mientras que no hay ítem con probabilidad nula de ser presentado. Una ecuación similar a las presentadas para el método progresivo y el método MRM permite empezar el test con selección aleatoria de items y, según el test avanza, ir

pasando a una selección de items paulatinamente más parecida a la de la optimización de la función de valoración (selección determinista).
Las simulaciones mostradas incluyen condiciones con bancos simulados y operativos, dos longitudes diferentes del test, presencia o ausencia del control de contenidos y restricciones o no en la tasa máxima de exposición. A través de esta variedad de escenarios, encontramos unas pautas comunes: la selección de items puede incoporar altos niveles de aleatoriedad, con lo que se mejora la seguridad, sin que con ello se produzcan pérdidas apreciables en la precisión de medida.

El método MRM y los métodos progresivo y proporcional, desde diferentes ángulos de la selección de items, convergen a resultados similares. De hecho, mientras que por un lado modificamos el modo de definir B_q y por otro las funciones de valoración de items, en todos los casos estamos aumentando la relevancia del azar en la administración de items. En el método de elegibilidad del ítem y en el método MRM, antes de la administración del test, se determina mediante experimentos aleatorios qué items van a componer B_q. Cuanto menor es el valor de la tasa máxima de exposición que queremos imponer, más items tienen parámetros de control de la exposición menores a 1, esto es, en más items interviene el azar como criterio de selección. El método MRM busca reducir al comienzo del test los valores de tasa máxima a sus mínimos posibles o valores próximos a estos. Por eso, el método MRM podría compartir también el título del estudio donde desarrollamos los métodos progresivo y proporcional: incorporando aleatoriedad a la función de información de Fisher. Líneas diferentes de considerar el problema de la seguridad del banco de items terminan por compartir una base común.

En el Capítulo 1, hemos descrito cuatro objetivos relevantes para los TAIs. Sin embargo, cuando se comparan diferentes reglas de selección, es habitual atenerse a únicamente dos de estos objetivos, precisión y seguridad. Así los hemos hecho en el estudio sobre el método MRM. En el estudio sobre los métodos progresivo y proporcional hemos evaluado, adicionalmente, el de facilidad de renovación del banco de items. Pese a simplificar la comparación de reglas al reducir el número de objetivos, el extraer conclusiones no resulta sencillo. Para que una regla pueda considerarse como superior a otra, tiene que ser mejor en un objetivo e igual o mejor en el otro. Sin embargo, esto no es lo que habitualmente se encuentra, puesto que es común encontrar un balance entre precisión y seguridad. Por ello, las conclusiones de los

estudios suelen incorporar un cierto grado de subjetividad. Tomemos como ejemplo la siguiente frase, del Capítulo 3:

> *"if an increment of 0.01 in RMSE is considered tolerable, the reduction in the overlap rate is 27%, with an acceleration parameter equal to 2"* (pág. 50).

Esta frase indica que la diferencia en precisión es tan pequeña que, de hecho, ambas reglas sí pueden considerarse como igualadas de selección en este objetivo. Una vez igualadas en un objetivo, fijando un criterio arbitrario de incremento no superior a 0.01, puede concluirse que una regla es mejor que la otra. Sin embargo, habría resultado posible reducir la tasa de solapamiento de la regla menos segura reduciendo el valor de tasa máxima impuesto. De este modo, podríamos haber llegado a igualar de hecho, y no aproximadamente, ambas reglas en una dimensión y compararlas en la otra. Conocemos que restricciones en la tasa máxima suponen reducciones en la precisión (p. ej., Barrada et al., 2008), pero podría ser que, para este caso, la reducción fuera incluso menor a 0.01. Ciertamente, una diferencia de 0.01 es pequeña, pero podríamcs estar encontrando diferencias entre métodos que se diluyeran al aplicar métodos más finos de comparación.

Vemos, por tanto, que los estudios comparativos entre reglas de selección, incluyendo los dos previamente comentados, adolecen de cierta arbitrariedad. Para intentar solucionarla, proponemos muestrear varios valores de tasa máxima de exposición, desde la mínima hasta la máxima posible. El modo de proceder viene descrito en el Capítulo 5, junto con los resultados para la comparación entre 6 reglas diferentes de selección.

Las conclusiones más importantes vienen comentadas al final de ese capítulo. Principalmente: (a) para test de 40 items o más, las diferencias entre reglas son escasas; (b) no hay ninguna regla que domine para todos los niveles de seguridad o de precisión, esto es, la regla a aplicar depende de los objetivos de la evaluación; y (c) para cualquier condición simulada, siempre hay una alternativa mejor a la regla más comúnmente empleada, la de máxima información puntual. Desde nuestro punto de vista, la aportación más importante es haber descrito el problema de la comparación entre reglas y haber propuesto un método sencillo y viable que permite la comparación. Consideramos que este procedimiento evitaría conclusiones tentativas y ayudaría a abandonar límites difusos de lo que es una pérdida de precisión de medida que se puede considerar como trivial como para considerar dos reglas iguales en precisión.

Hasta el momento, la literatura sobre TAIs ha compartido dos supuestos relativos a la seguridad del banco. El primero, que la distribución de las tasas de exposición de los items y la tasa de solapamiento son indicadores válidos de seguridad. El segundo, que la medida relevante de estos indicadores es la obtenida al final del test, esto es, que el solapamiento para posiciones diferentes al último ítem no aporta información con efectos aplicados.

En los tres primeros estudios presentados hemos mantenido estos supuestos. Así, por ejemplo, en el estudio segundo afirmábamos, en congruencia con el pensamiento generalizado en este área de investigación, "*the higher it is [the overlap rate], the greater the risk to security*" (pág. 54). Las variables básicas de comparación en seguridad entre los métodos MRM, proporcional y progresivo con respecto al método de máxima información puntual han sido la tasa de solapamiento y la distribución de tasas de exposición. Igualmente, las conclusiones del estudio tercero están basadas en la comparación en RMSE igualando la tasa de solapamiento. Sólo dando por válida la tasa de solapamiento como medida de seguridad tiene sentido el método de comparación entre reglas descrito.

Es común en los estudios sobre seguridad en TAIs informar únicamente de los resultados en precisión y seguridad conseguidos al final del test. Así lo hemos hecho nosotros en el estudio segundo y tercero. En el estudio primero, en la parte dedicada a la simulación mediante bancos generados aleatoriamente, ofrecimos los resultados para cada posición del ítem en el total del test, pero con intención de ilustrar cómo resulta posible aumentar el control de la exposición sin deteriorar importantemente la precisión. Es este mismo estudio, escribíamos que "*the relevant point in practical settings is what is obtained at the end of the test*" (pág. 46) o "*the exposure rates, overlap rates, RMSE, and bias at the end of the test are shown as these are the relevant data in a practical context*" (pág. 48).

El último estudio presentado puede considerarse una puesta en cuestión de estos supuestos. Hemos justificado tanto teórica como empíricamente que tasas de solapamiento mayores pueden dar lugar a bancos más seguros. Este resultado, inesperado desde la óptica de la investigación previa, puede explicarse por lo ocurrido en los items iniciales del test, datos habitualmente no disponibles, por considerarlos triviales.

Las conclusiones de este cuarto estudio tendrían implicaciones importantes. Si, como mantenemos, algunos de los supuestos básicos son discutibles, esto supone que también puede serlo la investigación que los ha dado por correctos. Por ejemplo, el

método alfa-estratificado (Chang & Ying, 1999) ha sido uno de los más investigados en los últimos años, puesto que parecía ser el que ofrecía mayor seguridad. Sin embargo, cuando evaluamos su funcionamiento en condiciones de filtrado de items, su ejecución no es la óptima.

Por último, este cuarto estudio pone de manifiesto la necesidad de explicitar todos los supuestos de investigación con los que se trabaja, para poder evaluar cuáles son apoyados por evidencias de diferente naturaleza, y si alguno es cuestionable. Al revisar los fundamentos, resulta posible desarrollar investigación con mayores garantías.

Resumiendo lo aportado por estos cuatro estudios en unos pocos puntos, lo central sería: (a) es posible incrementar la aleatoriedad en la selección de items sin deteriorar apenas la precisión; (b) la manipulación en múltiples niveles de tasa máxima permite una mejor comparación entre reglas de selección; (c) comenzar el test con selección aleatoria mejora la seguridad; y (d) el mejor modo de evaluar la seguridad del banco es simulando problemas en ésta.

7.2. Limitaciones

Los cuatro estudios presentados comparten algunas limitaciones. Entre ellas, el reducido número de condiciones diferentes simuladas. En el primer y segundo estudio se han puesto a prueba los métodos propuestos tanto con banco generados aleatoriamente como con un banco actualmente operativo. En los otros dos estudios únicamente se ha recurrido a bancos de ítems con parámetros generados de una distribución aleatoria. Actualmente, desconocemos si algún elemento de la composición de un banco de items, de la distribución de los niveles de rasgo de los examinados o de la longitud del test, por señalar tres aspectos claros en la configuración de un TAI, podría modular el patrón de resultados obtenidos. Pese a esto, entendemos que las aportaciones generales se mantendrían.

Otra limitación, común a casi toda la investigación sobre TAIs, es dar los parámetros de los items como carentes de error. Se asume que las estimaciones de los parámetros equivalen a sus valores reales. La incertidumbre sobre el valor real de los parámetros es pasada por alto, asumiendo que su efecto será irrelevante. Ésta es una línea de investigación todavía por cubrir.

Otro aspecto a destacar es que todos los resultados ofrecidos se obtienen en estudios de simulación. Asumimos que la regla de selección de items empleada no tendrá ningún efecto en el proceso de respuesta de los examinados. Hay algunas razones

para pensar que, tal vez, no sea así. Los métodos métodos MRM, progresivo proporcional suponen una elevada importancia del azar para la selección de items. Esto implica una alta oscilación entre las dificultades sucesivas de los items a administrar y que estos cambios no vienen marcados por la ejecución del examinado. Este efecto podría tener efectos sobre la motivación y el modo de hacer frente a la tarea. Sería, por tanto, altamente recomendable realizar estudios experimentales con aplicaciones reales de TAIs para evaluar si estos métodos con alto azar pueden conllevar algún efecto adverso.

Finalmente, mientras que el Capítulo 1 definíamos cuatro objetivos para un TAI, los resultados presentados se han ceñido, casi en su totalidad, a sólo dos, precisión y seguridad. Quedaría pendiente desarrollar métodos que hicieran posible la evaluación conjunta de todos los objetivos.

7.3. Futuras líneas de investigación

En los estos estudios presentados nos hemos centrado en bancos de items calibrados según el modelo logístico de tres parámetros y en tests de longitud fija. Una de las razones para esto, tal y como apuntábamos en el Capítulo 2, era la facilidad para extender las propuestas a modelos más recientes o complejos. Las alternativas más habituales a la configuración empleada aquí son los modelos politómicos; modelos multidimensionales o los tests de longitud variable. Una opción a considerar en el control de tasas máximas es el control condicionado a niveles de rasgo. La ampliación de las propuestas presentadas para incorporar estos diferentes modelos sería relativamente sencilla. Otras vías futuras de investigación pasan por resolver las limitaciones comentadas en el punto anterior.

7.4. Conclusiones

La visión clásica de la selección de items en TAIs ha sido buscar el ítem con mejores propiedades métricas según el nivel de rasgo del examinado. Estudios como los presentados demuestran que la optimización secuencial, para cada nuevo ítem a presentar, no ha de suponer los mejores resultados al final de test. Por ello, métodos que no buscan el mejor ítem para cada posición pueden llevar a resultados en precisión equivalentes a los que sí lo hacen. Hemos ilustrado esto mediante tres métodos diferentes, el MRM, el método progresivo y el proporcional.

Hemos ofrecido también métodos que permiten mejorar la comparación entre reglas de selección, tanto para el caso en el que se dan por válidas las variables

comúnmente empleadas para medir seguridad como en el caso de que se simule directamente filtrado de items.

De este modo, creemos haber ofrecido, por un lado, reglas que presentan un mejor funcionamiento que la regla de selección más común, y, por otro, modos nuevos de investigar en el campo de la seguridad en TAIs.

Referencias

(Los artículos incluidos en los Capítulos 3, 4, 5 y 6 contienen sus propias referencias al final. Sólo las citas procedentes del resto del texto se encuentran referenciadas en esta sección)

Ariel, A., Veldkamp, B. P., & van der Linden, W. J. (2004). Constructing rotating item pools for constrained adaptive testing. *Journal of Educational Measurement, 41*, 345-359.

Ban, J., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 191-212.

Ban, J., Hanson, B. A., Yi, Q., & Harris, D. J. (2002). Data sparseness and on-line pretest item calibration-scaling methods in CAT. *Journal of Educational Measurement, 39*, 207-218.

Barrada, J. R., Abad, F. J., & Olea, J. (2008, March). Varying the valuating function and the presentable bank in computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Barrada, J. R., Abad, F. J., & Veldkamp, B. P. (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema, 21*, 318-325.

Barrada. J. R., Mazuela, P., & Olea, J. (2006). Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema, 18*, 156-159.

Barrada, J. R., Olea, J., & Abad, F. J. (2008). Rotating item banks versus restriction of maximum exposure rates in computerized adaptive testing. *The Spanish Journal of Psychology, 11*, 618-625.

Barrada, J. R., Olea, J., & Ponsoda, V. (2007). Methods for restricting maximum exposure rate in computerized adaptative testing. *Methodology, 3*, 14-23.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2009). Item selection rules in computerized adaptive testing: Accuracy and security. *Methodology, 5*, 7-17.

Birnbaum, A. (1968). Some latent ability models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.) *Statistical theories of mental test scores* (pp. 392-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Buyske, S. (2005). Optimal design in educational testing. In M. P. F. Berger & W. K. Wong (Eds.), *Applied optimal designs* (pp. 1–16). New York: Wiley.

Chang, H. H.. (2004). Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In David Kaplan (Ed.) *The SAGE handbook of quantitative methodology for the social sciences* (pp. 117-133). Thousand Oaks, CA: Sage Publications.

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, H. H., & Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

Chang, H. H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73*, 441-450.

Chang, H.-H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics, 37*, 1466-1488.

Chang, H. H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67*, 387-398.

Chang, S. W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40*, 71-103.

Chang, S. W., & Harris, D. J. (2002, April). *Redeveloping the exposure control parameters of CAT items when a pool is modified*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans LA.

Chang, Y. C. I., & Lu, H. Y. (In press). Online calibration via variable length computerized adaptive testing. *Psychometrika*.

Chen, S. Y., & Ankenmann, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement, 41*, 149-174.

Chen, S. Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*, 241-255.

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*, 129-145.

Chen, S. Y., & Doong, S. H. (2008). Predicting item exposure parameters in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 61*, 75-91.

Chen, S. Y., & Lei, P. W. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement, 29*, 204–217.

Chen, S., Lei, P., & Liao, W. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 61*, 471-492.

Cheng, Y., & Chang. H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*, 369-383.

Cheng, Y., Chang, H. H., Douglas, J., & Guo, F. (2009). Constraint-weighted a-stratification for computerized adaptive testing with nonstatistical constraints. *Educational and Psychological Measurement, 69*, 35-49.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* New Jersey, Lawrence Erlbaum Associates.

Davey, T., & Parshall, C. G. (1995, April). New algorithms for item selection and exposure control with adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Davey, T., & Nering, N. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward, (Eds). *Computer-based testing: Building the foundation for future assesments* (pp. 165-191). Mahwah, NJ: Lawrence Erlbaum.

Dodd, B. G. (1990) The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14*, 355-366.

Eggen, T. J. H. M. (2001). *Overexposure and underexposure of items in computerized adaptive testing*. CITO, Measurement and Research Department Reports (2001-1).

Finkelman, M., Nering, M. L., & Roussos, L. A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement, 46*, 84-103.

Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8). Retrieved June 28, 2007, from http://www.jtla.org.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer.

Hau, K. T., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38*, 249-266.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington DC: American Psychological Association.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*, 241-261.

Lei, P., Chen, S., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43*, 245-264.

Leung, C. K., Chang, H. H., & Hau, K. T. (2005). Computerized adaptive testing: a mixture item selection approach for constrained situations. *British Journal of Mathematical and Statistical Psychology, 58*, 239-257.

Li, Y. H. & Schafer, W. D. (2003). The effect of item selection methods on the variability of CAT's ability estimates when item parameters are contaminated with measurement errors. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Li, Y. H., & Schafer, W. D. (2005). Increasing the homogeneity of CAT's item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. *Journal of Educational Measurement, 42*, 245-269.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95-100.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157-162.

Luecht, R. M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology, 7*(2). Retrieved July 24, 2009, from http://www.testpublishers.org/journal.htm.

Luecht, R. M., & Nungester R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 229-249.

Mills, C. N., & Steffen, M. (2000). The GRE computer adaptive test: Operation issues. In W. J. van der Linden and C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75-100). Boston: Kluwer Academic Press.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing, *Applied Measurement in Education, 9*, 287-304.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195.

Nering M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 115-127.

Olea, J., Abad, F. J., Ponsoda, V., & Ximénez, M. C. (2004). Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: Diseño y comprobaciones psicométricas. *Psicothema, 16*, 519-525.

Parshall, C. G., Spray, J. A., Kalohn, J. C., y Davey, T. (2002). *Practical considerations in computer-based testing*. Nueva York: Springer.

Rebollo, P., García-Cueto, E., Zardaín, P. C., Cuervo, J., Martínez, I., Alonso, J., Ferrer, M., & Muñiz, J. (2009). Desarrollo del CAT-Health, primer test adaptativo informatizado para la evaluación de la calidad de vida relacionada con la salud en España. *Medicina Clínica, 133*, 241-251.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Rubio, V., & Santacreu, J. (2003). *TRASI Test adaptativo informatizado para la evaluación del razonamiento secuencial y la inducción como factores de la habilidad intelectual general*. Madrid: TEA ediciones.

Rulison, K. L., & Loken. E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33*, 83-101.

Segall, D. O. (2004). A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 29*, 439-460.

Stern, E. B. & Havlick, L. (1986). Academic misconduct: Results of faculty and undergraduate student surveys. *Journal of Allied Health, 15*, 139-142.

Stocking, M. L., & Lewis, C. L. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57-75.

Stocking, M. L., & Lewis, C. L. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice* (pp. 163-182). Dordrecht, The Netherlands: Kluwer Academic.

Stocking, M. L, & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measuremen, 69*, 778-793.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201-216.

van der Linden, W.J. (2000). Constrained adaptive testing with shadow tests. In W.J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27-52). Boston, MA: Kluwer Academic Publishers.

van der Linden. W. J. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 28*, 249-265.

van der Linden, W. J. (2005). A comparison of item-selection method for adaptive tests with content constraints. *Journal of Educational Measurement, 45*, 283-302.

van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 13*, 35-53.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259-270.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273-291.

van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics 32,* 398-418.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203-226.

Wainer, H. (2000a). CATs: Whither and whence. *Psicológica, 21*, 121-133.

Wainer, H. (2000b). Rescuing computerized testing by breaking Zipf's law. *Journal of Educational and Behavioral Statistics, 25*, 203-224.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17-27.

Zwick, R. (2000). The assessment of differential item functioning in computer adaptive tests. In W. van der Linden & C.A. W. Glas (Eds.), *Computerized Adaptive Testing. Theory and Practice* (pp. 221-243). Boston, MA: Kluwer Academic Publishers.

REUNIDO EN EL DÍA DE LA FECHA, EL TRIBUNAL QUE SUSCRIBE ACORDÓ CONCEDER A LA PRESENTE TESIS DOCTORAL LA CALIFICACIÓN DE Sobresaliente "Cum laude"

MADRID, 20 - 1 - 2010

EL PRESIDENTE, EL SECRETARIO,

FDO.: RAFAEL SAN MIGUEL FDO.: EDUARDO DOVAL

PRIMER VOCAL SEGUNDO VOCAL TERCER VOCAL

FDO.: PEDRO MONTALVO FDO.: MIGUEL A. GARCIA FDO.: Carmen García.